مدخل إلى اللسانيات الحاسوبية

أ. د. منصور بن محمد الغامدي
د. عبدالعزيز بن عبدالله المهيوبي
أ. أحمد روبي محمد عبدالرحمن
د. إشراق علي أحمد الرفاعي
د. صلاح راشد الناجم
د. وليد بن عبدالله الصانع

تحرير: د. عبدالله بن يحيى الفيفي

مباحث لغوية ٣٠

مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية
King Abdullah Bin Abdulaziz Int'l Center for The Arabic Language

www.kaica.org.sa

مباحث لغوية ٣٠

مدخل إلى اللسانيات الحاسوبية

تأليف:

أ. د. منصور بن محمد الغامدي
د. عبدالعزيز بن عبدالله المهيوبي
أ. أحمد روبي محمد عبدالرحمن
د. إشراق علي أحمد الرفاعي
د. صلاح راشد الناجم
د. وليد بن عبدالله الصانع

تحرير :

د. عبدالله بن يحيى الفيفي

مركز الملك عبدالله بن عبدالعزيز الدولي
لخدمة اللغة العربية
King Abdullah Bin Abdulaziz Int'l Center for
The Arabic Language

مركز الملك عبدالله بن عبدالعزيز الدولي
لخدمة اللغة العربية
King Abdullah Bin Abdulaziz Int'l Center for
The Arabic Language

مدخل إلى اللسانيات الحاسوبية

الطبعة الأولى
١٤٣٨ هـ - ٢٠١٧م

المملكة العربية السعودية - الرياض
ص.ب ١٢٥٠٠ الرياض ١١٤٧٣
هاتف:٠٠٩٦٦١١٢٥٨٧٢٦٨ - ٠٠٩٦٦١١٢٥٨١٠٨٢
البريد الإليكتروني: nashr@kaica.org.sa

ح/ مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية، ١٤٣٨هـ.
فهرسة مكتبة الملك فهد الوطنية أثناء النشر
الفيفي ، عبدالله بن يحيى
مدخل إلى اللسانيات الحاسوبية./ عبدالله بن يحيى الفيفي
- الرياض، ١٤٣٨هـ
..ص؛ .. سم
ردمك: ٠-٢-٩٠٨٩٥-٦٠٣-٩٧٨
١- اللغة العربية-معالجة البيانات أ. العنوان
ديوي ٤١٨,٢٨٥ ١٤٣٨/٤١٨٣
رقم الإيداع: ١٤٣٨/٤١٨٣
ردمك: ٠-٢-٩٠٨٩٥-٦٠٣-٩٧٨

التصميم والإخراج
دار وجوه للنشر والتوزيع
Wojooh Publishing & Distribution House
www.wojoooh.com
المملكة العربية السعودية - الرياض
الهاتف:4562410 الفاكس:4561675
للتواصل والنشر:
info@wojoooh.com

بسم الله الرحمن الرحيم

مقدمة المحرر [1]

الحمد لله رب العالمين، والصلاة والسلام على أشرف المرسلين، نبينا محمد وعلى آله وصحابته أجمعين، وبعد:

يعد مجال اللسانيات الحاسوبية (Computational Linguistics) أحد العلوم البينية (Interdisciplinary) التي تقع بين علمين مستقلين، وذلك لاتصاله بعلم اللغويات أو اللسانيات من جهة، وبعلم الحاسب الآلي من جهة أخرى. ويرى مارتن كي (Kay, ٢٠٠٣)[2] أن اللسانيات الحاسوبية قد برزت إلى حيز الوجود خلسة وبخجل، وأن بدايتها كانت في عام ١٩٤٩م عندما كتب وارن ويفر مذكرته الشهيرة التي يشير فيها إلى إمكانية بناء نظام للترجمة الآلية. ثم تلا ذلك عقد أول مؤتمر للترجمة الآلية في معهد

١ – عبدالله بن يحيى الفيفي: أستاذ اللغويات الحاسوبية المساعد في جامعة الإمام محمد بن سعود الإسلامية في الرياض. درس البكالوريوس في اللغة العربية في جامعة الملك خالد في أبها، والماجستير في تعليم اللغة بمساعدة الحاسب في قسم اللغويات في جامعة Essex، والدكتوراه في اللغويات الحاسوبية في قسم الحاسب الآلي في جامعة Leeds، وكلاهما في بريطانيا. له عدة أبحاث منشورة حول تقنيات معالجة اللغة العربية آلياً، والمدونات اللغوية وبرامجها الحاسوبية، وكذلك مدونات المتعلمين، والمعاجم الحاسوبية لمتعلمي اللغة العربية، إضافة إلى مشاركته في تأليف بعض الكتب المتخصصة في اللسانيات الحاسوبية، والمدونات اللغوية وتطبيقاتها. عمل محكماً لدى عدد من الدوريات العلمية والمؤتمرات الدولية. أنشأ المدونة اللغوية لمتعلمي اللغة العربية، وأسس فريق معجم المفردات الشائعة لمتعلمي اللغة العربية "شائع".

2- Kay, Martin (2003) Introduction. In: Mitkov, Ruslan (Ed.), The Oxford Handbook of Computational Lnguistics. New York: Oxford University Press.

ماساتشوستس للتكنولوجيا (MIT) في ١٩٥٢م، ثم صدرت مجلة علمية بعنوان الترجمة الآلية في ١٩٥٤م. أما مصطلح اللسانيات الحاسوبية نفسه فقد بدأ استعماله في منتصف الستينات (1960s)، ويُرجح أن ديفيد هيز (David Hays) هو أول من أطلق هذا المسمى على هذا المجال عندما كان عضواً في اللجنة الاستشارية لمعالجة اللغة آلياً في الأكاديمية الوطنية للعلوم في الولايات المتحدة الأمريكية. ثم أتى بعد ذلك عدد من المتخصصين الذين كان لهم دور في ظهور هذا المجال مثل نعوم تشومسكي (Noam Chomsky) وجون كوك (John Cocke)، وغيرهم. واليوم يعد هذا التخصص من التخصصات ذات الأهمية المتزايدة لما له من دور كبير في التطور الحاصل في مجال معالجة اللغة الطبيعية (Natural Language Processing) والذكاء الاصطناعي (Artificial Intelligence)، إذ بات أحد الأسس المهمة في صناعة التطبيقات الحاسوبية المستعملة في كثير من مظاهر الحياة اليومية.

ويُعرف نيوقس اللسانيات الحاسوبية بأنها فرع عن علمي اللغة والحاسب، يهدف إلى تصميم نماذج رياضية للتراكيب اللغوية؛ للتمكن من معالجة اللغة آلياً عن طريق الحاسب، كما يعرفه من وجهة نظر لغوية على أنه تشكيل للنظريات والنماذج اللغوية أو تنفيذها على الآلة، ويرى أنه بإمكاننا النظر إليه على أنه وسيلة لتطوير نظريات لغوية جديدة بمساعدة الحاسب (Nugues, 2006) [1].

ولقد شهد البحث في مجال اللسانيات الحاسوبية تقدماً متسارعاً في السنوات القليلة الماضية، مما ساعد على بروز تطبيقات عملية استفادت من نتائج تلك الأبحاث بشكل مباشر وفي مجالات شتى، لعل من أبرزها تطبيقات التخاطب مع الآلة المسماة بتطبيقات المساعد الشخصي الذكي (Intelligent personal assistant) والتي نرى انتشارها بين أيدينا مثل: سيري (Siri) من شركة أبل (Apple)، وجوجل ناو (Google Now) من شركة جوجل (Google)، وكورتانا (Cortana) من شركة مايكروسوفت (Microsoft)، وأمازون إيكو (Amazon Echo) من شركة أمازون (Amazon)، وعشرات الأنظمة المشابهة التي تجمع عدداً من مستويات المعالجة اللغوية في تطبيق واحد. ومن هنا تبرز أهمية وجود مدخل إلى اللسانيات الحاسوبية باللغة العربية،

1- Nugues, Pierre M. (2006) An Introduction to Language Processing with Perl and Prolog. Berlin: Springer-Verlag.

لتعريف القارئ العربي بهذا المجال وببعض فروعه وتطبيقاته، وليكون تمهيداً لما يكتب بعده من مراجع متخصصة تتناول فروعه بتوسع أكثر. وهذا الكتاب موجه بالدرجة الأولى لطلاب الدراسات العليا في الجامعات، أو الراغبين في الاطلاع على هذا المجال من غير المتخصصين، إذ يقدم تعريفاً لعدد من مجالات اللسانيات الحاسوبية وهي:

- الصوتيات الحاسوبية Computational Phonetics.
- التحليل الصرفي Morphological Analysis.
- التحليل النحوي Syntactic Parsing.
- التحليل الدلالي Semantic Analysis.
- تحليل النصوص Texts Analytics.
- التدقيق الإملائي Spelling Checker.

ولقد حرص المشاركون في تأليف هذا الكتاب على أن يكون الطرح تعليمياً متدرجاً مع شرح المصطلحات قدر الإمكان، وتقريب المعلومات للقارئ بأمثلة واضحة تساعد على الفهم والتطبيق. وفيما يلي عرض موجز لمحتويات الكتاب اعتماداً على الملخصات التي سترد لاحقاً في بداية كل فصل من فصوله.

ففي **الفصل الأول** يتحدث منصور الغامدي عن **الصوتيات الحاسوبية**، مبتدئاً بمقدمة عامة لعلم الصوتيات، ثم يتطرق إلى الفروع الثلاثة لهذا العلم: الصوتيات النطقية، الصوتيات الأكوستية، الصوتيات السمعية. وتحت كل فرع يُورد مقدمة ثم يذكر التقنيات المتعلقة به من حيث الدراسة والبحث والتحليل. ولأن هذا العلم أساس لعدد من العلوم، فالفصل يذكر التطبيقات التقنية لعلم الصوتيات وخاصة في مجال التعرف الآلي على الكلام وتوليد الكلام آليا، والتعرف على المتحدث، مع الإشارة إلى المتطلبات التي تقوم عليها هذه التقنيات.

في **الفصل الثاني** يتحدث عبدالعزيز المهيوبي عن **التحليل الصرفي**، مبتدئاً بعرض موجز لخصائص الصرف العربي، ثم مفهوم التحليل الصرفي الآلي، وقواعد المعطيات المصاحبة للمحلل الصرفي. ينتقل بعد ذلك إلى الحديث عن مجموعة من الأسس المهمة لبناء محلل صرفي دقيق للغة العربية، ويقدّم نظرة تاريخية للتحليل الصرفي الآلي، مع استعراض مجموعة من أهم المحللات الصرفية العربية، مشيراً لأهمية التطبيقات الحاسوبية للتحليل الصرفي. ينتقل بعد ذلك إلى الحديث عن مجموعة من الضوابط

والمحددات التي تساعد في بناء المحللات الصرفية، مقسِّماً إيّاها إلى ضوابط ومحددات شكليَّة ودلالية. ثم يستعرض المشكلات التي تواجه بناء محلل صرفي دقيق لكلمات اللغة العربية ونصوصها، وطرق عرض نتائجها، وكيفية توصيف القواعد الصرفية لبناء المحلل الصرفي الآلي. ثم يشير في عجالة إلى أسباب قصور المحللات الإنجليزية عن استيعاب خصائص اللغة العربية، متحدثاً بالتفصيل عن خطوات بناء المحلل الصرفي الآلي، ومتطلبات بنائه.

وفي **الفصل الثالث** يتحدث أحمد روبي عن **التحليل النحوي**، فيقدم رؤية شاملة عن التحليل النحوي الحاسوبي في إطار تطبيقي، محاولاً الوقوف على منطلقات التحليل النحوي (التمثيل النحوي - النظرية النحوية - المحتوى النحوي) وأدواته في صورة مبسطة، بحيث تكون مدخلا مبسطًا للقارئ العربي، يمكن من خلالها فهم الصورة العامة لإطار التحليل النحوي الحاسوبي. وسعيًا لتحقيق هذه الغاية، فإنه يقف على قوام العملية النحوية/ التركيبية ودورها في بناء التطبيقات الحاسوبية المختلفة التي تناظر الأداء الإنساني؛ فيأتي الفصل في خمسة محاور رئيسية: تتضمن مقدمة يعرض من خلالها تأصيل طرق التوصيف النحوي، ثم عرضًا لإرهاصات التحليل النحوي الحاسوبي، ثم أهمية التحليل النحوي الحاسوبي للدراسات اللغوية بصورة عامة وللغويات الحاسوبية أو معالجة اللغة الطبيعية بصورة خاصة، ويلي ذلك الخطوات الإجرائية اللازمة لبناء أية عملية تحليل نحوي حاسوبي، والتي يمكن تلخيصها في العناصر التالية على الترتيب: (النص الخام/ المدونة اللغوية - تجزئة النصوص - العنونة بالأجزاء الكلامية - الترميز بالعلاقات التركيبية)، وأخيرا يعرض الفصل بعض موارد التحليل النحوي المتاحة للغة العربية وكذلك تطبيقاته.

في **الفصل الرابع** تتحدث إشراق الرفاعي عن **التحليل الدلالي**، إذ يقدم الفصل نبذة تعريفية عن التحليل الدلالي، تشمل استعراضا لأهم المصطلحات المرتبطة بهذا المجال مثل المتضادات والمترادفات، إضافة للفرق بين المعنى الحرفي والمعنى العملي للنصوص، كما يشير الفصل إلى المنهج البحثي المستخدم في دراسة التحليل الدلالي، إضافة إلى أبرز الموارد اللغوية المتاحة مثل شبكة الكلمات العربية (Arabic WordNet). يتطرق الفصل فيما يلي ذلك إلى عدد من أبرز تطبيقات هذا المجال وهي: تحليل العواطف، وفك اللبس الدلالي، مع تعريف كل منهما، واستعراض أبرز ما أنجز فيهما من أبحاث

وتطبيقات. كما يتطرق الفصل إلى الحديث عن الكينونات كمفهوم مهم عند دراسة التحليل الدلالي، ويُعنى بدراسة الكلمات المجردة والعلاقات فيما بينها من حيث المعنى. يشير الجزء الأخير من الفصل إلى أبرز الجهود البحثية في مجال التحليل الدلالي فيما يخص اللغة العربية، والتي قُدمت من قبل مجموعات بحثية شهيرة حول العالم؛ حتى يتسنى للقارئ الاطلاع على المخرجات البحثية الأحدث في هذا المجال.

في **الفصل الخامس** يتحدث صلاح الناجم عن **تحليل النصوص**، فيتناول أهمية تحليل النصوص كتطبيق أساسي من تطبيقات المعالجة الحاسوبية للغة الطبيعية، وهو يساعد على اكتشاف وانتزاع معرفة هامة من نصوص حرة لا تسير وفق بنية منظمة (Unstructured Data). يشير الفصل كذلك إلى التطور الكبير في مجال البيانات الضخمة (Big Data) الذي أفرز كميات هائلة من البيانات النصية، ومنها على سبيل المثال لا الحصر مشاركات وحوارات وسائل التواصل الاجتماعي، إذ يتطلب تحليل هذه البيانات إيجاد تطبيقات ومنصات تحليلية ولغات برمجة وأدوات وخوارزميات (Algorithms) متخصصة للتعامل مع هذا الكم الهائل من البيانات النصية. وهنا تبرز أهمية تحليل النصوص كمجال بيني (Interdisciplinary) يدمج أكثر من مجال أكاديمي أهمها علم الحاسوب، اللسانيات الحاسوبية، استرجاع المعلومات (Information Retrieval)، تحليل البيانات (Data Mining)، تعلُّم الآلة (Machine Learning)، والإحصاء (Statistics). يتحدث الفصل أيضاً عن أهمية البيانات الضخمة، ومستويات ومراحل تحليل النصوص، ثم ينتقل إلى الحديث عن المعالجة الحاسوبية للنصوص وخطواتها، ثم يتناول بعض التطبيقات مثل تصنيف النصوص، وانتزاع المعلومات، وتحليل المزاج العام.

وفي **الفصل السادس** يتحدث وليد الصانع عن **التدقيق الإملائي**، فيستعرض أبرز التحديات التي تواجه مطوري المدققات الإملائية للغة العربية، إذ تعتبر اللغة العربية من اللغات المدعومة في كثير من أنظمة التشغيل وأجهزة الحاسب الآلي والبرمجيات، وقد قامت كبريات الشركات العالمية بتطوير مدققات إملائية للغة العربية. ويعد تطوير مدققات إملائية عربية تحديا يواجه مطوري هذا النوع من التطبيقات بسبب اختلاف صيغ الإملاء زماناً ومكاناً.

يُعرّج الفصل بعد ذلك على آليات اكتشاف الأخطاء الإملائية وإشكالياتها، ومن

ثم أبرز الطرق لتصحيح هذه الأخطاء الإملائية، كما يعطي نبذة سريعة عن بعض النظريات المتقدمة التي تستخدم في أبحاث تطوير المدققات الإملائية، وبعض المراجع الأساسية التي قد تفيد القارئ.

ختاماً، أتقدم بالشكر الوافر – بعد شكر الله عز وجل – إلى القائمين على مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية، على ما قدموا لفريق التأليف من دعم متصل وتذليل للعقبات، في سبيل خروج هذا الكتاب في أتم صورة، ليكون مرجعاً للمهتمين بهذا الميدان، وبداية للسالكين فيه من طلاب الجامعات وخصوصاً طلاب الدراسات العليا. كما أتقدم بالشكر الجزيل لجميع المشاركين في تأليف فصول هذا الكتاب الذين قدموا خلاصة فكرهم، مع ما أبدوا من التزام ودقة في العمل، كل ذلك في سبيل تحقيق غاية الكتاب، وانتظام عقد موضوعاته.

المحرر/ عبدالله بن يحيى الفيفي

الرياض
١٧ ربيع الثاني ١٤٣٨هـ
ayjfaifi@gmail.com

◇◇◇

الفصل الأول

الصوتيات الحاسوبية

أ. د. منصور بن محمد الغامدي[1]

ملخص البحث

يقدم هذا الفصل مقدمة عامة لعلم الصوتيات تاريخا وتطورا عبر الزمن. ثم يتطرق إلى فروع هذا العلم الثلاثة: الصوتيات النطقية، الصوتيات الأكوستية، الصوتيات السمعية. وفي كل فرع يقدم له ثم يذكر التقنيات المتعلقة به من حيث الدراسة والبحث والتحليل. ولأن هذا العلم علم أساس لعدد من العلوم، وحيث إن الكتاب عن التقنيات، فالفصل يذكر التطبيقات التقنية لعلم الصوتيات وخاصة في مجال التعرف الآلي على الكلام وتوليد الكلام آليا والتعرف على المتحدث. والمتطلبات التي تحتاج لها هذه التقنيات من علم الصوتيات، حيث تشكل الصوتيات أساسا لها كعلم وتطبيق وإجازة لما ينتج من تقنيات، إضافة إلى مدى دقتها وملاءمتها للمستفيدين.

١- حاصل على درجة الدكتوراه في الصوتيات. له أكثر من ثمانين كتابا وبحثاً علمياً منشوراً. حاصل على خمس براءات اختراع. أشرف وشارك في أكثر من عشرين مشروعاً بحثياً، نتج عنها نظم حاسوبية وبرمجيات وخوارزميات وقواعد بيانات. أشرف على رسائل دكتوراه. حاضر في جهات متخصصة عن تطبيقات الصوتيات كحوسبة اللغة، وعيوب التخاطب، والترجمة، واكتساب اللغة الأم، وتعلم اللغة الأجنبية. حكم أعمالا بحثية وإبداعية وبحوثا مقدمة للنشر. شارك في عدد من الهيئات واللجان. عمل في قطاعات الدولة أربعاً وأربعين سنة؛ تقلد خلالها عددا من المهام. حصلت بحوثه المنشورة على أكثر من خمسمئة استشهاد على موقع «قوقل سكولار». (m.ghamdi2@Qiyas.org)

الفصل الأول: الصوتيات الحاسوبية

١ . المقدمة

الصوتيات أحد فروع علوم اللسانيات ويشكل المستوى الأدنى من مستويات الدراسات اللغوية. حيث تكون المستويات الأعلى (علوم الدلالة والنحو والصرف والمعاجم) عقلية مجردة بينما الصوتيات علم ملموس. فهو يتعلق بأصوات اللغة من حيث مخارجها وخصائصها الأكوستية وسماعها. وله ارتباط مباشر بعلم لساني آخر أعلى منه وهو الفونولوجيا phonology الذي يشمل دراسة النظام الصوتي للغة وعلاقة الأصوات ببعضها وتأثير بعضها على بعض.

فالفونولوجيا يحدد الأصوات الأساسية لكل لغة أو ما يعرف بالفونيمات phonemes (جمع فونيم phoneme)[١]. وتتشكل فونيمات اللغة ما يعرف بالنظام الصوتي sound system. ويتكون النظام الصوتي من صوامت consonants و صوائت vowels. فنظام العربية الصوتي يحتوي على ٢٨ صامتا (الجدول: ١) و ٦ صوائت هي: الفتحة

١ – الفونيم هو أصغر وحدة صوتية تغير المعنى في كلمات اللغة. فالصوتان / س/ و / ص/ فونيمان مختلفان في العربية كما في كلمتي "سد" بمعنى حاجز الماء، و "صد" بمعنى أعرض. وهما ليسا كذلك في الإنجليزية، فلو نطقا في كلمة مثل seed "بذرة" لما غيرت في معناها. والصوتان / p/ و / b/ في الإنجليزية فونيمان مختلفان في كلمات مثل: pat "تربيتة" و bat "خفاش" بينما ليسا كذلك في العربية، فلو نطقا في كلمة مثل "بات" بمعنى أصبح، لما غيرت في معناها.

القصيرة Ʌ (a) والطويلة Ʌ́ : (a:)، والضمة القصيرة V́ (u) والطويلة V́ : (u:)، والكسرة القصيرة V (i) والطويلة V́ : (u:) (1). وفي المستوى الفونولوجي تُطبَّق القوانين الفونولوجية فتتأثر بعض الأصوات نتيجة لوجودها في بيئة صوتية محددة كما في حال إدغام assimilation اللام الشمسية في الصوت الذي يليها عند وجودها قبل الأصوات: / ث، ن، ذ، ظ، ت، د، ط، ض، ر، ل، س، ز، ص، ش/ ، كما في "الثّابت" و "السّابق" (الغامدي وآخرون، ١٤٢٤هـ). تتم العمليات الفونولوجية كغيرها من العمليات في مستويات اللغة العليا في الدماغ البشري، ومع إرسال الإشارات العصبية من الدماغ إلى الجهاز الصوتي يبدأ المستوى الصوتي phonetic level. حيث يمكن قياس الإشارة العصبية الواردة من الدماغ إلى عضلات الجهاز الصوتي (2) vocal tract. ثم متابعة حركة الجهاز الصوتي والأصوات اللغوية التي يخرجها.

	Bilabial شفتاني	Labiodental شفوي أسناني	Interental بين أسناني	Alveodental لثوي أسناني	Alveopalatal غاري لثوي	Palatal غاري	Velar طبقي	Lab-velar شفوي طبقي	Uvular لهوي	Pharyngeal حلقي	Glottal حنجري
Nasal أنفي	m م			n ن							
Stop شديد	b ب			t ت d د			k ك		q ق		ʔ ء
Emphatic Stop*				tˤ ط dˤ ض							
Fricative رخو		f ف	θ ث ð ذ	s س z ز	ʃ ش				χ خ ʁ غ	ħ ح ʕ ع	h هـ
Emphatic fricative**			ðˤ ظ	sˤ ص							
Affricate مزجي					dʒ ج						
Glide ليني						j ي		w و			
Lateral جانبي				l ل							
Trill تكراري				r ر							

الجدول ١ : نظام العربية الصوتي (الصوامت). الصفوف الرأسية تعبر عن مخرج الصوت والأفقية عن كيفية خروجه. *شديد مفخم، **رخو مفخم (الغامدي، ١٤٣٦هـ).

يقدم هذا الفصل معلومات عامة عن علم الصوتيات والتقنيات المستخدمة لدراسته، إذ هي مفتاح للحصول على معلومات دقيقة عنه، والتي بدورها تشكل

١- الرموز المستخدمة هنا حسب الألفبائية الصوتية العربية والألفبائية الصوتية الدولية (بين قوسين)، لمزيد عن هذه الرموز (الغامدي، ١٤٢٧هـ، أ)

٢- في الجهاز الصوتي ما يقرب من مائة عضلة.

أساس حوسبة الصوتيات أو تطبيقاته الحاسوبية. بعد ذلك يتطرق الفصل إلى ثلاثة من أبرز التطبيقات في مجال حوسبة الصوتيات وهي: التعرف الآلي على الكلام، وتوليد الكلام آليا، والتعرف على المتحدث آليا. هذه التطبيقات أصبحت مؤخرا ملموسة في حياة الناس اليومية خاصة الأول والثاني.

ولما يتسم به العصر الحالي من تقدم في تقنية المعلومات information technology وشيوع تطبيقاتها applications بين أفراد المجتمع على الإنترنت والحاسبات والأجهزة الكفية حتى لم يعد للإنسان غنى عنها، فهي الرابط بينه وبين الآخرين وبينه وبين التطور المتسارع للمعرفة، فإن هذه التقنية قائمة على التطور المذهل الذي حدث في السنوات الأخيرة في اللسانيات الحاسوبية computational linguistics على جميع مستوياتها من الدلالة والبراغماتية إلى الفونولوجيا والصوتيات. فأصبح بالإمكان الكتابة بجميع اللغات وكذلك عرض رموزها الكتابية وطباعتها. كما أصبح بإمكان الحاسبات معالجة نصوصها وحفظها واستعادتها وفهرستها، بل وأبعد من ذلك الترجمة من لغة إلى أخرى وفهم النصوص وتلخيصها وأيضا توليد نصوص جديدة لمواضيع محددة.

ومن الجدير بالذكر أن اللغة العربية من اللغات القلائل التي رموز كتابتها فونيمي؛ أي أن لكل فونيم رمز كتابي "قرافيم" خاص به. فالصوت /ف/ يكتب دائما هكذا "ف". هذه السمة قليلة الحدوث في اللغات الأخرى، فعلى سبيل المثال، فونيم اللغة الإنجليزية /f/ يظهر في الكتابة بعدة أشكال: «ph», «gh», «f» كما في الكلمات: fast, laugh, physics تباعا. هذه السمة تساعد في أمور شتى من تعلم القراءة والكتابة وحتى توليد الكلام والتعرف عليه آليا إضافة إلى معالجة النصوص وتحليلها.

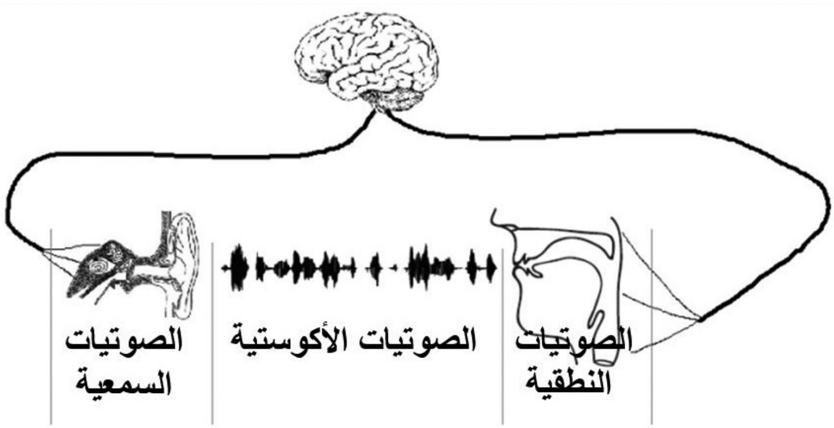
٢. الصوتيات

عرفت الدراسات الصوتية منذ القدم، ولكن أبرز دراسة عميقة ومفصلة كانت على يد الخليل بن أحمد الفراهيدي وتلميذه سيبويه (أبو بشر عمرو بن عثمان بن قنبر) في القرن الثاني الهجري. ومن أبرز ما وصل إلينا منها كتاب «الكتاب» لسيبويه الذي قدم من خلاله وصفا دقيقا لأصوات اللغة العربية وقواعدها الفونولوجية [1]. ولم تكن هناك

١- تندر الإشارة للجهود العربية في مجال الصوتيات في المراجع الغربية، كما هي الحال لجهود العرب في العلوم الأخرى التي ليس لها مكان في المراجع الغربية، رغم حضورها الملموس في مفردات اللغات الأوربية كدليل واضح على أن الحضارة الغربية قامت على الحضارة العربية. من المراجع البارزة في تحليل كتاب سيبويه فيما يتعلق بالصوتيات والفونولوجيا ما كتبه الدكتور عبدالمنعم الناصر في رسالته للدكتوراه (Al-Nassir, 1985).

إضافة تذكر من بعد ذلك إلا بعد الثورة الصناعية في أوربا في القرن الثاني عشر الهجري (أي بعد ألف سنة) حيث بدأت النهضة الأوربية وازدهرت معها كافة العلوم بما فيها الصوتيات، على سبيل المثال الهنغاري «وولفغانغ» Wolfgang von Kempelen الذي صنع أول آلة نطق (Ohala, 1991). واستمر تطور علم الصوتيات كغيره من العلوم في العصر الحديث ليتأسس على قواعد علمية صلبة فيما يتعلق بجمع قواعد البيانات ذات العلاقة بأصوات اللغة أو الدراسات والاستنتاجات مما مهد لتطبيقات عملية في حياة الناس كاكتساب أصوات اللغة (بالنسبة للأطفال كلغة أم، وللكبار كلغة أجنبية)، وعلاج عيوب التخاطب، واختبارات اللغة، والتعرف على المتحدث، والتواصل مع الآلة صوتيا، والتعرف الآلي على الكلام، وتوليد الكلام آليا.

ينقسم علم الصوتيات إلى ثلاثة فروع هي: الصوتيات النطقية والصوتيات الأكوستية والصوتيات السمعية (الشكل ١). ويتضح من اسم كل منها المجالات التي تعنى بها. ولعلنا نستعرض هذه الفروع بشيء من التفصيل.

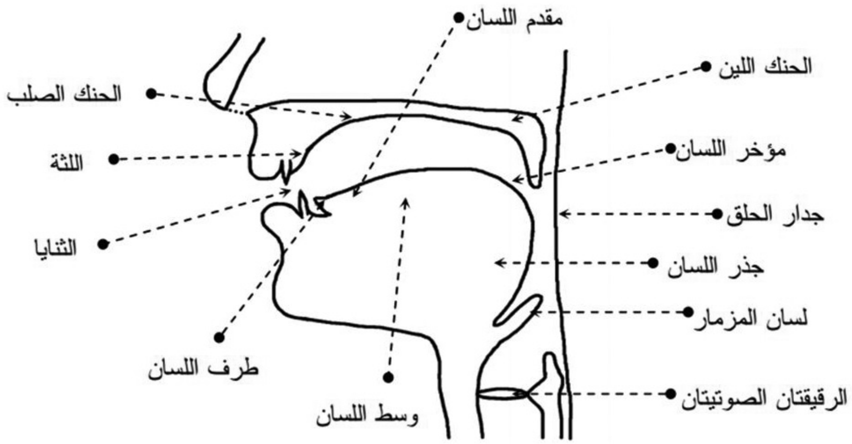


الشكل ١: فروع الصوتيات الثلاثة.

٢, ١ الصوتيات النطقية

يتابع علم الصوتيات النطقية articulatory phonetics الإشارة العصبية بعد صدورها من الدماغ متجهة إلى عضلات الجهاز الصوتي التي تقدر بمائة عضلة. ويمكن معرفة وصول الإشارات العصبية إلى عضلات جهاز النطق باستخدام تقنية

Electromyography والمعروفة اختصارا EMG حيث تدخل إبرة بها مجس electrode في العضلة لتحديد وقت تأثرها بالإشارة العصبية مما يسهل معرفة علاقة ذلك بإخراج صوت قيد الدراسة. ويتكون الجهاز الصوتي vocal tract كما في الشكل ٢ من عدد من أعضاء النطق speech organs التي تتحكم فيها مجموعة من العضلات بناء على ما يصلها من إشارات عصبية قادمة من الدماغ. هذه الأعضاء تتحكم في ثلاثة تجاويف (الحلقي والفموي والأنفي) حيث تتحكم في شكل التجويفين الأولين وفي مخارج الهواء عبر التجاويف جميعها. ويتسبب التغيير في أشكال تجاويف النطق في إخراج أصوات مختلفة ومتعددة تتجاوز المائتي صوت يستخدمها الإنسان للتخاطب مع الآخرين عبر اللغات البشرية المختلفة.



الشكل ٢: أعضاء النطق في الجهاز الصوتي (الغامدي، ١٤٣٦هـ).

اعتمدت الدراسات السابقة للجهاز الصوتي على الأشعة السينية لدراسة حركة أعضائه أثناء الكلام وتحديد حجم التجاويف، أما الآن فإن magnetic resonance imaging أو ما تعرف اختصارا MRI تقدم بيانات أكثر دقة ومزيدا من التفاصيل ومرونة في استخدام الحاسب لقياس الأبعاد والمساحات والأحجام كما في الشكل ٣ (Sorensen,et al. 2016).



الشكل ٣: صورتان للرأس أخذت بجهاز MRI. الأولى على اليمين تظهر فيها اللهاة قد فتحت مجرى الهواء للخروج عبر التجويف الأنفي بينما تظهر الصورة اليسرى انغلاق مجرى الهواء (Nakai, et al. 2016).

ويوظف الجهاز الصوتي الهواء في توليد الأصوات بطرق متعددة، أكثرها استخداما هو هواء الزفير؛ حيث تعترض أعضاء النطق الهواء الخارج من الرئتين مسببة خروج الأصوات المطلوبة لكل لغة. ويوصف الصوت اللغوي حسب نطقه بمكان خروجه في الجهاز الصوتي place of articulation وكيفية إخراجه manner of articulation (الجدول ١).

الجهاز الصوتي موحد من حيث وظيفته وتشريحه لكل بني البشر رغم الاختلاف في الشكل والحجم من شخص إلى آخر ومن مجموعة إلى أخرى، إلا أن هذه الاختلافات لا تؤثر في مجمل نطق أصوات اللغة، فجهاز الإنسان الصوتي قادر على نطق أصوات أية لغة. إلا أن بعض الأجهزة الصوتية يتسبب حجم وشكل أعضائها في منح الكلام سمة أكوستية. فصغر الرقيقتين الصوتيتين، على سبيل المثال، عند النساء والأطفال تجعل ترددهما عالٍ وهذا من أسباب تمييزنا لأصوت الأطفال والنساء والرجال.

يضم الجهاز الصوتي الأعضاء المشار إليها في الشكل ٢، حيث تعترض هواء الزفير الخارج من القصبة الهوائية مولدة بذلك أصوتا عديدة. وأول هذه الأعضاء اعتراضا للهواء هما الرقيقتان الصوتيتان vocal folds المحميتان بصندوق غضروفي يسمى الحنجرة larynx، حيث تتذبذبان بترددات مختلفة حسب طولهما، فيتميز الرجال

بتردد منخفض (١٠٠ مرة في الثانية تقريبا)[١] والنساء بتردد عال (٢٠٠ مرة في الثانية تقريبا) والأطفال بتردد أعلى (٤٠٠ مرة في الثانية تقريبا). وتؤثر طبيعة الكلام على تردد الرقيقتين الصوتيتين، فيزيد ترددهما عند التحدث بصوت عالٍ وينخفض عند التحدث بصوت منخفض للمتحدث نفسه، وتتميز الجمل الخبرية بانحدار ترددهما بينما الجمل الاستفهامية بتصاعد التردد. ويخرج من بين الرقيقتين الصوتيتين صوتان في العربية هما / ء/ ، / هـ/ . يلي الحنجرة من الأعلى التجويف الحلقي pharyngeal cavity حيث يمتد من الحنجرة إلى اللهاة في الأعلى. وفي التجويف الحلقي لسان المزمار epiglottis الذي يخرج منه الصوتان / ح/ ، / ع/ عندما يلتقي بالجدار الحلقي pharyngeal wall. يجد الهواء عند خروجه من التجويف الحلقي مساران: الأول يؤدي إلى التجويف الأنفي nasal cavity والآخر إلى التجويف الفموي oral cavity. التجويف الأنفي هو التجويف الوحيد الثابت في حجمه وطوله وذلك لعدم وجود أعضاء متحركة داخلة ولثباته داخل إطار من العظام والغضاريف. تعمل اللهاة كبوابة للتجويف الأنفي، حيث تغلقه إذا ارتفعت فتمنع الهواء من المرور فيه، وتفتحه إذا انخفضت فيخرج الهواء منه. ويخرج من الأنف صوتان في العربية هما / ن/ ، / م/ . أما التجويف الفموي فأكثر التجاويف مرونة ولهذا تخرج معظم الأصوات منه. ومن أبرز أعضائه اللسان lingua /tongue الذي يرمز في كثير من اللغات إلى اللغة ﴿ وَهَٰذَا لِسَانٌ عَرَبِيٌّ مُّبِينٌ ﴾[سورة النحل آية ١٠٣]، ﴿ وَمِنْ ءَايَٰتِهِۦ خَلْقُ ٱلسَّمَٰوَٰتِ وَٱلْأَرْضِ وَٱخْتِلَٰفُ أَلْسِنَتِكُمْ وَأَلْوَٰنِكُمْ ۚ إِنَّ فِى ذَٰلِكَ لَءَايَٰتٍ لِّلْعَٰلِمِينَ ﴾ [سورة الروم آية ٢٢]. ويخرج من اللسان وما يقابله من سقف الفم الأصوات الآتية: اللهاة uvula: / خ/ ، / غ/ ، / ق/ ؛ الحنك اللين velum إضافة إلى الشفتين lips: / و/ ؛ الحنك اللين: / ك/ ؛ الحنك الصلب hard palate: / ي/ ؛ بين الحنك الصلب واللثة alveolar ridge: / ش/ ، / ج/ ؛ اللثة والأسنان tooth: / ن/ ، / ت/ ، / د/ ، / ط/ ، / ض/ ، / س/ ، / ز/ ، / ص/ ، / ل/ ، / ر/ ؛ بين الأسنان: / ث/ ، / ذ/ ، / ظ/ . العضوان الآخران المتحركان في التجويف الفموي هما الشفتان حيث تشكلان البوابة الخارجية للتجويف الفموي. ويخرج بينهما الصوتان / م/ ، / ب/ . وهناك صوت يخرج نتيجة التقاء الشفة السفلى مع الثنايا العليا وهو / ف/ .

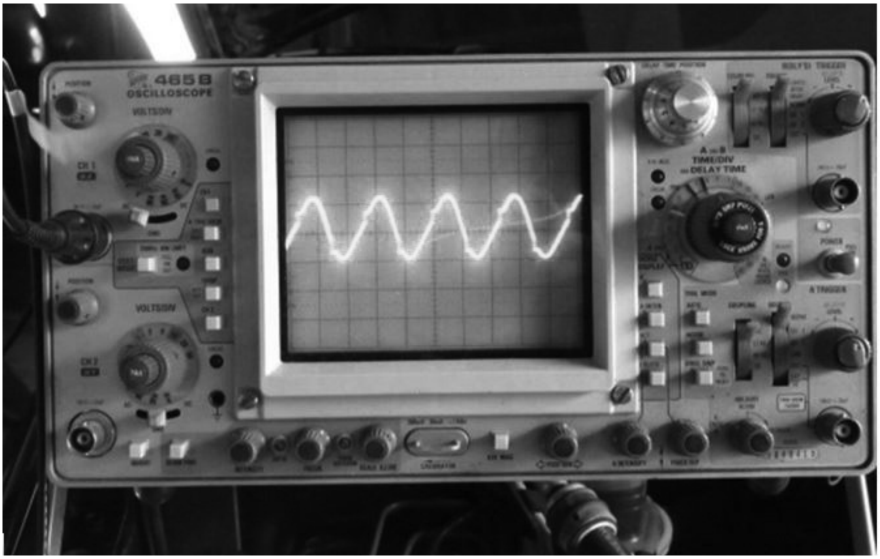
١ - تعرف الدورة الكاملة للموجة الصوتية بالهيرتز Hertz

مما يلاحظ خروج مجموعة من الأصوات من مخرج واحد، فكيف تكون مختلفة؟ طريقة إخراج الصوت تجعل له خصائص أكوستية مختلفة. فمثلا، / ن/ ، / ت/ ، / د/ ، / ط/ ، لها المخرج نفسه ولكن: / ن/ يخرج من الأنف، / ت/ يخرج من الفم وهو مهموس وغير مطبق، / د/ يخرج من الفم وهو مجهور وغير مطبق، / ط/ يخرج من الفم وهو مهموس ومطبق.

يشكل تردد الرقيقتين الصوتيتين التردد الأساس للكلام fundamental frequency وهو منخفض نسبيا (١٠٠-٤٠٠ هيرتز، يتفاوت حسب المتحدث وطبيعة الكلام)، إلا أن التجاويف التي تعلو الحنجرة (الحلق والفم والأنف) تنتج رنينا resonance داخل التجويف مولدا ما يعرف بالتوافقيات harmonics وهي تكرار منتظم لتردد الرقيقتين الصوتيتين قد يتجاوز العشرة آلاف هيرتز (Auditory Neuroscience). وتختلف التوافقيات من حيث الشدة amplitude فالأماكن التي تكون فيها الشدة عالية تسمى نطق رنينية formants حيث تلعب دورا مهما في التفريق بين الصوائت.

٢ , ٢ الصوتيات الأكوستية

يخرج الصوت من الجهاز الصوتي على هيئة موجات صوتية تنتشر في الوسط المحيط بالمتحدث لتصل إلى إذن السامع. ويسمى العلم المختص بالموجات الصوتية للكلام بالصوتيات الأكوستية acoustic phonetics. ولأن الموجات الصوتية لا تشاهد بالعين، كان من الصعب دراستها علميا حتى ظهر الأوسلوسكوب oscilloscope في نهاية القرن التاسع عشر (الشكل ٤). وتعرض شاشة الأوسلوسكوب ترددات الموجات الصوتية ببعدين: الزمن والتردد. فيمكن حساب تردد الرقيقتين الصوتيتين في زمن محدد كما يمكن معرفة نوع الموجة الصوتية هل هي بسيطة كالصادرة عن الشوكة الرنانة أم مركبة كالصادرة عن الجهاز الصوتي.



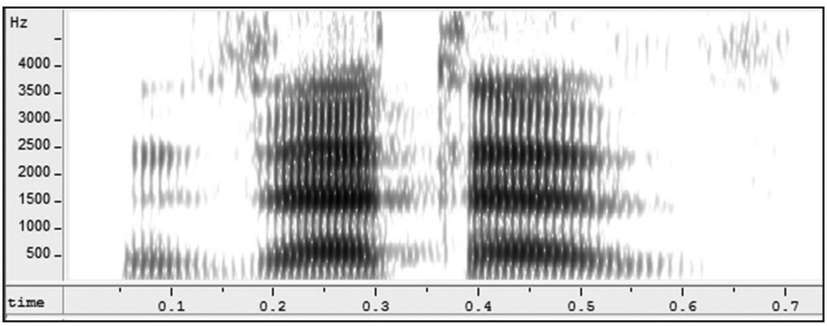
الشكل ٤: جهاز أوسلوسكوب (Wikimedia).

ولأن جهاز الأوسلوسكوب لا يعرض تفاصيل دقيقة عن الموجة الصوتية، ظهر في منتصف القرن العشرين جهاز المطياف spectrograph، حيث أمكن رؤية الإشارة الصوتية بثلاثة أبعاد: الزمن والتردد وشدة كل تردد مما سهل على الباحثين والمطورين دراسة الأصوات اللغوية وتطوير النظم الإليكترونية والحاسوبية ذات العلاقة بها (الشكل ٥).



الشكل ٥: جهاز المطياف (Universiteit Leiden).

ويعرض جهاز المطياف رسما طيفيا للموجة الصوتي spectrogram يبين المحور الأفقي فيه (الزمن)، والرأسي (التردد). كما يقدم الرسم الطيفي معلومات قيمة تحويها الموجة الصوتية منها: تردد الرقيقتين الصوتيتين (الخطوط الرأسية)، وترددات النطق الرنينية (الخطوط الأفقية)، وشدة الموجة (السواد). ولا يزال الرسم الطيفي يستخدم إلى الآن في الدراسات الصوتية وكثير من التطبيقات منها التعرف على هوية المتحدث.



الشكل ٦: رسم طيفي لموجة صوتية، المحور الأفقي للزمن والرأسي للتردد.

وظلت الرسوم الطيفية الناتجة عن استخدام جهاز المطياف[1] أساس الدراسات الصوتية الأكوستية إلى أواخر القرن العشرين، إذ مع تطور الإلكترونيات والبرمجيات أصبح جهاز المطياف الذي كان يحتاج لغرفة مجرد برمجيات على الحاسب أو تطبيق على الأجهزة الكفية. ويبين الجدول ٢ بعض برمجيات تحليل الموجات الصوتية المفتوحة التي يستخدمها مؤخرا دارسو موجات الكلام وكذلك مطورو النظم الحاسوبية ذات العلاقة بها، حيث يستطيع الباحث تحميلها والاستفادة منها مباشرة، أو الدخول على شفرة البرنامج لتطويره أو التعديل عليه لخدمة أهداف الدراسة والبحث. كما أن هناك نظم حاسوبية أخرى ليست مجانية من أشهرها MATLAB الذي له تطبيقات واسعة في مجالات الهندسة والدراسات والتحليل والاستنتاج، وكذلك Computerized Speech Lab ويتكون من برمجيات software إضافة إلى عتاد hardware وهو مخصص لتسجيل وتحليل الموجات الصوتية الخاصة بالكلام.

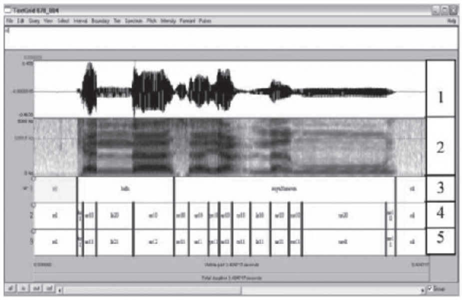
الاسم	الجهة المنتجة	الرابط الإلكتروني
PRAAT	University of Amsterdam	http://www.fon.hum.uva.nl/praat/
WaveSurfer	Royal Institute of Technology	http://www.speech.kth.se/wavesurfer/
WASP	London Global University	http://www.phon.ucl.ac.uk/resource/sfs/wasp.php
SIL Speech Analyzer	SIL Interna-tional	http://www-01.sil.org/computing/sa/index.htm?_ga=GA1.2.1982728125.1471423724

الجدول ٢: اسماء ومواقع تحميل بعض برمجيات تحليل الإشارة الصوتية الشائعة والمفتوحة.

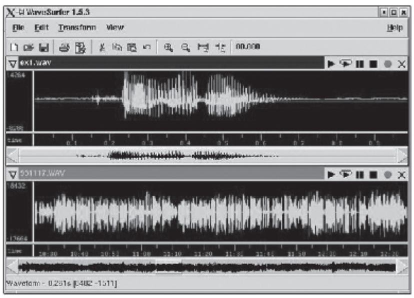
غالبا ما تعرض برمجيات تحليل الإشارة الصوتية الموجة على شكلين: موجة صوتية waveform ورسم طيفي spectrogram إضافة إلى تفاصيل دقيقة إما على شكل رسوم

١- تطبع على ورق باستخدام إبرة كهربائية تسخن بناء على شدة الصوت فتعلم على الورق منتجة رسما طيفيا.

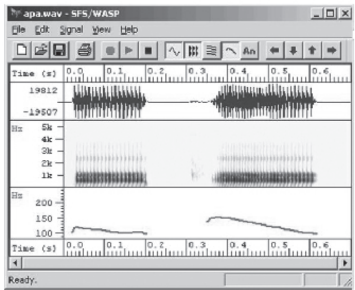
بيانية أو رقمية لخصائص موجة صوتية في فترة محددة، حيث يمكن تحديد موجة صوت لغوي ليعرض البرنامج خصائص ذلك الصوت وما يحويه من ترددات ونطق رنينية formants وتردد أساس fundamental frequency وغيرها. وتمكن هذه البرامج الدارس من التعديل على خصائصها الأكوستية كالحذف والإضافة والتقطيع والترميز. وقد سهلت هذه الخصائص على الدارسين معرفة الكثير عن الإشارة الصوتية وما تحمله من خصائص وأسرار كالمشعرات الصوتية acoustic cues التي بموجبها يستطيع الانسان التعرف على الأصوات والتمييز بينها كما في حالة التمييز بين الصوت المهموس والمجهور وكذلك الربط بين الصوت وناطقه (Singh, et al. 2016).



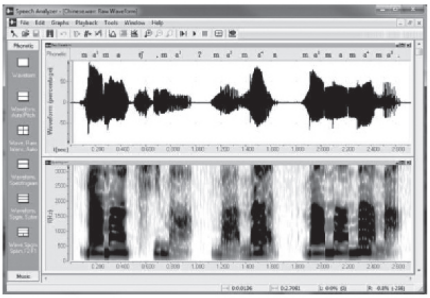
واجهة استخدام PRAAT



واجهة استخدام WaveSurfer



واجهة استخدام WASP



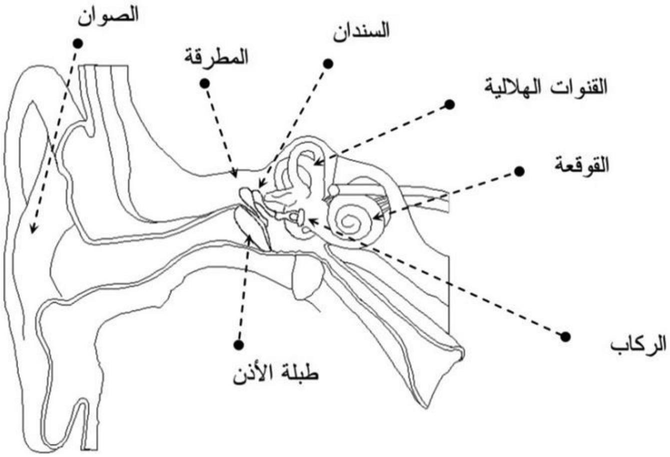
واجهة استخدام SIL

الشكل ٧: واجهات بعض برمجيات تحليل الموجات الصوتية الشائعة والمفتوحة.

وتشكل الصوتيات الأكوستية أساسا مهما للتطبيقات التقنية ذات العلاقة بالكلام البشري كالتخاطب عن بعد (نظم الاتصالات كالهاتف والاتصالات اللاسلكية) التي تعتمد على الخصائص الفيزيائية للكلام لنقل كلام مفهوم وواضح للمستخدم بأقل التكاليف في استخدام الطاقة والتطوير والصناعة التقنية. وكذلك في تطوير نظم حاسوبية معقدة للتعرف الآلي على الكلام وتوليده آليا والتعرف على المتحدث.

٢ , ٣ الصوتيات السمعية

يبدأ عمل هذا التخصص من ملامسة الموجات الصوتية للأذن الخارجية إلى تعرف الدماغ على الأصوات والتمييز بينها. وتشكل الأذن العضو الأساس في هذا العلم (الشكل ٧). وتقوم بعض أعضاء السمع كالقناة السمعية والعظيمات في الأذن الوسطى بتضخيم ترددات محددة (لها علاقة مباشرة بموجات الكلام) عشرات المرات مما يسهل على الإنسان التعرف على الكلام عن طريق الموجات الصوتية التي يسمعها. ويمكن للأذن البشرية سماع الموجات التي يقع ترددها بين ٢٠ هيرتز و ٢٠ كيلوهرتز وهو نطاق أعلى بكثير من تردد موجات الكلام الصوتية التي تقع بين ١٠٠ هيرتز و١٠ كيلوهرتز تقريبا.



الشكل ٨: الأذن البشرية وما تحويه من أعضاء سمعية (الغامدي، ١٤٣٦هـ).

استفاد الباحثون من خصائص الأذن البشرية لتطوير تقنيات تلبي حاجة الإنسان كأنظمة تكبير الصوت في الأماكن العامة والتخاطب عن بعد والتسجيل الصوتي والاتصالات، بحيث يكون الصوت - خاصة ما يتعلق بالكلام - واضحا بما فيه الكفاية ليفهم السامع ما نطقه المتحدث. وتقوم هذه التقنيات على دراسات صوتية للكلام تبين الموجات المهمة للسامع والأخرى الأقل أهمية. فمثلا نظم الاتصالات لا تنقل جميع ترددات الكلام بين المتحدث والسامع، ذلك لأن نقلها جميعا مكلف عليها من حيث الطاقة والتقنية. لهذا تقتصر على الترددات الأقل من أربعة آلاف هيرتز، وتحذف الترددات الأعلى منها. هذا النطاق كاف للإنسان لفهم الكلام ولو أن جودة

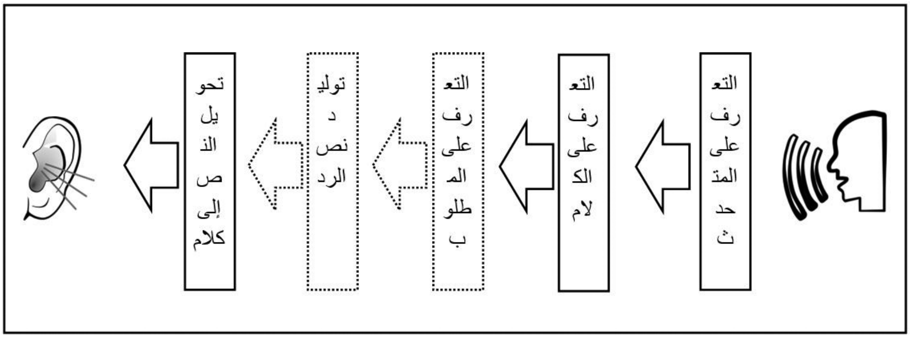
الموجة الصوتية ليست كما هي في حال سماع المتحدث مباشرة. ولكنها توفر على نظم الاتصالات أكثر من ٦٠٪ من تكلفة موجات صوت الكلام الكاملة.

٣. تطبيقات وتقنيات ذات علاقة بالصوتيات

أسهمت التقنيات الحديثة في تطوير تطبيقات عديدة ذات علاقة بالكلام، فكانت البداية عندما تمكن الإنسان من تحويل طاقة الموجة الصوتية إلى أنواع أخرى من الطاقة؛ ميكانيكية في البداية كما في الفونوقراف phonograph (المعروف في السعودية "بكم")، ثم إلى طاقة كهربائية (Robjohns, 2001). حيث صنع لاقطا صوتيا (ميكرفون) يحول الموجة الصوتية التي تصل إليه إلى طاقة كهربائية. كانت هذه البداية في غاية الأهمية إذ بناء عليها تمكن الإنسان من تطوير تقنيات أكثر تعقيدا. فطور نظم الاتصالات (موجة صوتية (المتحدث) > طاقة كهربائية (الهاتف) > موجة صوتية (المستمع)). وتمكن بهذه التقنية أيضا من حفظ الموجة الصوتية ليستعيدها فيما بعد وقت ما شاء (موجة صوتية > طاقة كهربائية > طاقة مغناطيسية > طبعة مغناطيسية يمكن حفظها). كما أنه تمكن أيضا من بث موجات الكلام الصوتية عبر الأثير إلى الناس ليلتقطوها بجهاز المذياع الذي يعيدها إلى موجات صوتية يمكن سماعها. كانت هذه التقنيات بداية مهمة لأعمال قادمة، حيث تمكن الإنسان مع نهاية القرن العشرين من التحول إلى التقنية الرقمية حيث تحول الموجة الصوتية إلى أرقام يمكن التعامل معها بسهولة في الحفظ أو الإرسال والاستقبال أو التحليل والتشفير.

هناك ثلاث تقنيات من أكثر التقنيات ذات العلاقة بالكلام التي تشكل تحديا كبيرا أمام الباحثين والمطورين وذات أهمية كبيرة للمستخدمين بجميع شرائحهم، هي: توليد الكلام آليا speech synthesis أو text-to-speech، التعرف الآلي على الكلام automatic speech recognition أو speech-to-text، التعرف على المتحدث speaker recognition أو speaker identification. هذه التقنيات الثلاث مهمة في التعامل مع الآلة، فقد كان حلم الإنسان أن يتمكن من التخاطب مع الجماد وها هو الحلم قد اقترب كثيرا بل وتحقق إلى درجة شبه مقبولة. هذه التقنيات الثلاث مترابطة إذ تشكل أساس التخاطب مع الآلة. فالتعرف على المتحدث هو البصمة أو المفتاح الذي يوصل المتحدث إلى بياناته ويسمح له بتنفيذ أعماله مستخدما صوته، والتعرف

على الكلام وسيلة لإيصال الأوامر والطلبات وإدخال المعلومات للآلة، وتوليد الكلام وسيلة لاستجابة الآلة للإنسان صوتياً. الأنموذج في الشكل ٨ مثاليٌّ للتخاطب مع الآلة؛ فهو مكون من خمسة نظم حاسوبية، إلا أنه قد لا تتوفر جميع هذه النظم فيكتفى بواحد منها. فعلى سبيل المثال فتح الباب بالبصمة الصوتية لا يتطلب إلا نظاماً واحداً (التعرف على المتحدث). وكذلك استخدام الكفيف لنظام قارئ الكتاب الآلي لا يتطلب سوى نظام الناطق الآلي.



الشكل ٩: أنموذج متكامل للتخاطب مع الآلة. الإطار المتصل للنظم المشار إليها في هذا الفصل، الإطار المتقطع لنظامي حاسب آخرين.

كما أن تحويل الموجة الصوتية إلى كهرباء كانت قفزة في تاريخ التقنيات فإن تحويل الموجات الصوتية إلى حروف وكلمات يعتبر قفزة في تقنية البرمجيات. لأنه لا يمكن التعامل مع موجة أصوات الكلام من الناحية اللغوية، ولكن يمكن التعامل مع النص بطرق عدة منها: الترجمة الآلية فالترجمة الآلية لا تتم مباشرة من موجة صوتية بلغة ما إلى موجة صوتية للغة أخرى، بل هي: موجة صوتية > نصوص > ترجمة آلية > نصوص > موجة صوتية. وكذلك العمليات المعقدة الأخرى ذات العلاقة باللغة كفهم اللغة والتنقيب في النصوص وتلخيصها وفهرستها وتصنيفها وتقويمها وتوليدها آليا، كل هذه العمليات وغيرها الكثير تتطلب توفر نصوص لغوية. لهذا فإن التعرف الآلي على الكلام في غاية الأهمية ليس فقط لتطبيقات مباشرة ولكن لأنه يمثل بنية أساسية لعمليات أخرى أكثر تعقيداً.

تحويل الموجة الصوتية إلى كهرباء تتطلب مهندساً كهربائياً ليقوم بالمهمة، أما تحويل الموجة الصوتية إلى نص فيتطلب متخصصين من علوم شتى (مبرمجين، مطورين،

لغويين، أصواتيين) إضافة إلى قواعد بيانات وأدوات حاسوبية. يضاف إلى ذلك التعقيدات المصاحبة لموجات أصوات الكلام التي من أهمها تباينها من متحدث إلى آخر بل للمتحدث الواحد. فكل مرة ينطق شخص كلمة محددة تكون لها موجة صوتية مختلفة مهما تكرر نطقها. عدم الثبات هذا يشكل تحديا للمبرمجين والمطورين. ورغم التقدم الذي وصلت إليه تقنيات التعامل مع الكلام، إلا أن الطريق لا يزال طويلا للوصول إلى نظم حاسوبية يكون أداؤها قريبا من أداء الإنسان حتى تكون طبيعية مقبولة من مستخدميها. ورغم ذلك فقد ظهرت نظم مقبولة إلى حد بعيد من المستخدمين لعل من أشهرها في الوقت الحاضر نظام «سيري»(Siri Wikipedia) الذي يأتي محملا على أجهزة أبل الكفية الآن وهو نظام للتخاطب مع الجهاز يقوم بالوظائف التالية:

التعرف الآلي على الكلام بتحويل الموجة الصوتية إلى نص.

معالجة النص وتحليله في محاولة لفهم المطلوب منه، تحديد الرد على المطلوب في النص إما بتنفيذه كأوامر، أو البحث في ذخيرة الهاتف أو في الإنترنت لإيجاد مقاربة للمطلوب والعودة بنتائج المقاربة.

تنفيذ أوامر محددة كالاتصال بشخص معلوماته متاحة على الهاتف، أو تحويل نصوص نتيجة المقاربة إلى كلام.

ويعمل نظام "سيري" حاليا بعشرين لغة منها العربية، وبعضها بأكثر من لكنة كالإنجليزية التي تحمل تسع لكنات بما فيها الأسترالية والأمريكية والكندية والبريطانية.

٣,١ التعرف الآلي على الكلام

التعرف الآلي على الكلام هو تطوير نظام حاسوبي يتولى تحويل الموجة الصوتية إلى نص. فالموجة الصوتية المعروضة في الشكل ٨ هي لـ «كلا سيعلمون» (متجهة من اليسار لليمين) وكما يلاحظ فالموجة الصوتية متصلة ببعضها دون انقطاع سواء بين الأصوات أو بين الكلمتين. فلكي يتعرف الحاسب على هذه الموجة، عليه أولاً: معالجة الإشارة الصوتية ليتمكن من التعامل معها، ثانياً: تحديد الفواصل/ الحدود بين كل صوت وآخر على الموجة الصوتية، ثالثاً: استخلاص الخصائص الأكوستية للأصوات بين الفواصل، رابعاً: مقارنتها بالخصائص الصوتية المخزنة لديه، خامساً: اختيار المقارب لكل صوت حسب أعلى احتمالية ممكنة ليصل إلى نتيجة أن هذه الموجة مكونة من الأصوات الآتية: ك ʌ ل ل ʌ: س ʌ ي ʌ ع ل ʌ م ʋُ: ن، ٦) تحويل هذه الرموز إلى

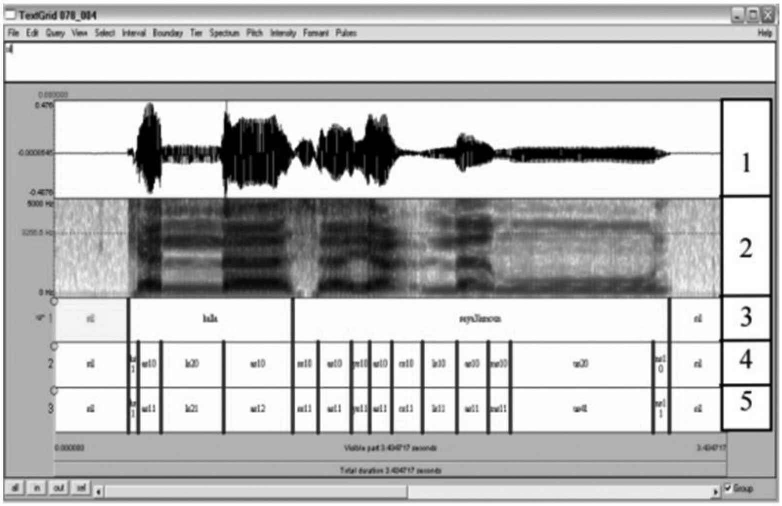
النص المقروء "كلا سيعلمون". عندها تكون مهمة نظام التعرف الآلي على الكلام قد انتهت. إلا أن كل مرحلة من هذه المراحل تشكل تحديا كبيرا للمتخصصين من مبرمجين وباحثين ومطورين.

ولتنفيذ المراحل السابقة فإن نظام التعرف على الكلام المطلوب تطويره يحتاج إلى:

(١) قاعدة بيانات صوتية لكلام عدد من المتحدثين تكون موجاته الصوتية مقطعة segmented ومرمَّزة transcribed كما في الشكل ١٠ . وهناك عدد من قواعد البيانات للغة العربية منها LDC و WestPoint Arabic وبنك الأصوات الهاتفية لمتحدثين سعوديين (Alotaibi, et al. 2012).

(٢) استخدام أدوات حاسوبية كأدوات أنموذج ماركوف المخفي (Hidden Markov Model Toolkits HTK) تقوم ببناء أنموذج أكوستي acoustic mode من قاعدة البيانات باستخلاص السمات الأكوستية لكل صوت sound features، وتقوم لاحقا بمقارنة ما يصلها من كلام بما هو مخزن لديها في الأنموذج الأكوستي للوصول إلى النص المقابل للموجة الصوتية.

(٣) معالجة النص المستخلص ليكون نصا لغويا سليما.



الشكل ١٠: موجة صوتية للكلمتين "كلا سيعلمون" (Alghamdi, et al. 2007).

يستخدم الباحثون والمطورون عددا من الآليات والأدوات للرفع من كفاءة نظم التعرف التي تقاس باستخدام ملفات صوتية مخصصة لاختبار النظام، فتحسب نسبة

عدد الفونيمات التي تعرف عليها النظام إلى إجمالي عدد الفونيمات في ملفات الاختبار. وقد وصلت في بعض البحوث المنشورة مؤخرا إلى دقة ٨٩٪ (Nahar, et al. 2016).

يواجه مطورو أنظمة التعرف الآلي على الكلام عدداً من التحديات من أهمها التباين في أصوات المتحدثين واختلاف صوت المتحدث الواحد من وقت لآخر. كما أن الخلفية الصوتية للمتحدث قد تتسبب في إرباك النظام بسبب عدم قدرته على فصل الموجات الصوتية الصادرة من المتحدث من الأصوات المنتشرة في الخلفية. كما يصعب على هذه الأنظمة معرفة ما وراء الكلام كحالة المتحدث النفسية والعاطفية وما يريد فعلا من كلامه في حالات مثل: "نعم" التي قد يقصد بها المتحدث الإجابة على سؤال أو التعجب أو الإنكار. فهذه الإمكانات لا يزال ينتظرها المزيد من البحث والتطوير لإضافتها لأنظمة التعرف الآلي على الكلام.

٣, ٢ توليد الكلام آليا

تهدف هذه الأنظمة إلى تحويل النص المكتوب إلى موجات كلام مسموعة. وغالبا ما تقوّم هذه الموجات بناء على: (١) الوضوح intelligibility /comprehensibility، وهو مدى قدرة الإنسان على فهم كلام النظام بمعنى تحويل الموجات الصوتية إلى النص الذي نطقه النظام في الأصل (Chang, 2011). (٢) الطبيعية naturalness وهي مدى قربها من الكلام الطبيعي الذي يولده جهاز صوت الإنسان. هذا التقويم يضعه مطورو هذه الأنظمة نصب أعينهم عند العمل على تطوير نظام لتوليد الكلام آليا. كما أن المستخدمين يأخذون هذين المعيارين في الحسبان عند الشراء أو الرغبة في استخدام نظام من هذا النوع.

هناك عدد من الطرق لتوليد الكلام آليا لعل من أكثرها شيوعا: (١) توليف الموجات المجزأة concatenated speech synthesis، وتعتمد هذه الطريقة على استخلاص وحدة صوتية speech unit لكل صوت يمثل فونيما في اللغة من كلام طبيعي، وعند توليد الكلام تستدعى الوحدات الصوتية التي تقابل النص المطلوب نطقه، وتدمج مع بعضها لتمثل صوتا متصلا. (٢) التوليد الآلي للكلام باستخدام النطق الرنيني formant based synthesis، وهذه الطريقة لا تحتاج إلى قواعد بيانات صوتية، وإنما توليد النطق الرنيني الذي يقابل كل فونيم، وتدمج مع بعضها لتولد الموجة الصوتية

المطلوبة للنص. (٣) التوليد الآلي للكلام بناء على قواعد بيانات صوتية لأحد المتحدثين data driven synthesis، ويستخدم المطورون لهذه الطريقة أدوات حاسوبية كأنموذج ماركوف المخفي، والشبكات العصبية (neural network Wu, et al. 2016). وقد شاع انتشار الطريقة الأخيرة نظرا للنتائج التي تعطيها من حيث قربها من الصوت الطبيعي. وتعتمد هذه الطريقة على قواعد بيانات صوتية سجلت ورمزت بعناية لأحد المتحدثين (Almosallam, et al. 2013).

٣,٣ التعرف على المتحدث آليا

تحمل الموجات الصوتية الصادرة عن الجهاز الصوتي مشعرات خاصة بالأصوات اللغوية (الفونيمات) وتحمل أيضا معلومات أخرى غير لغوية منها حالة المتحدث النفسية/ المزاجية (سعيد، حزين، غضبان، ...)، وكذلك البصمة الخاصة به، فالمستمع قادر على التمييز بين المتحدثين، فيتعرف على المتحدث المعتاد على سماع صوته إذا كان ضمن متحدثين آخرين، كما أنه يميز المتحدثين البالغين من صغار السن، وكذلك صوت الرجل من الأنثى. ولوجود تطبيقات عديدة لخاصية التعرف على المتحدث، ظهرت محاولات لأتمتتها ليستفيد منها الإنسان في حياته اليومية.

كغيرها من النظم الحاسوبية ذات العلاقة بالكلام البشري، فإن أنظمة التعرف على المتحدث تحتاج إلى قواعد بيانات صوتية لمتحدثين. كما أنها تحتاج لأدوات حاسوبية لبناء النظام ومن أشهر هذه الأدوات أنموذج خليط غاسيون(Gaussian mixture model Islam, et al. 2016). حيث يقوم الأنموذج ببناء أنموذج أكوستي خاص بكل متحدث في قاعدة البيانات الصوتية لكي يتعرف عليه من خلال صوته عندما يعرض عليه صوت جديد لنفس المتحدث. وكما في نظم التعرف على الكلام، فإن قواعد البيانات الصوتية تنقسم إلى قسمين: (١) مجموعة التدريب training set، وهي الجزء من القاعدة المستخدم في تدريب النظام لبناء أنموذجه الأكوستي ويشكل عادة ٩٠٪ من القاعدة. (٢) مجموعة الاختبار testing set، وهي مجموعة الملفات من القاعدة المستخدمة في تقويم النظام. وتعتمد كفاءة النظام على نسبة التعرف على أصوات المتحدثين في مجموعة الاختبار. وينشر المطورون نتائج بحوثهم بعرض نسب التعرف التي وصلوا إليها، فهي معيار جودة ما توصلوا إليه.

ومن المنتجات المتاحة في أسواق البرمجيات ذات العلاقة بالتعرف على المتحدث: Open Sesame، Nuance VocalPassword، Authentify، VoiceVault، iAM BioValidation، VoiceBiometrics Group، Voice Print System. حيث تستخدم هذه الأنظمة في تطبيقات شتى فهي تعمل عمل المفتاح الذي يمكن الدخول به على الحاسبات الشخصية أو الهواتف الذكية أو حساب على الإنترنت أو فتح أبواب البيت أو الغرف وما شابه ذلك. هذه الأنظمة تستجيب فقط لصاحب الصوت المبرمج على ذلك.

وللتعرف على المتحدث تطبيقات أخرى لها علاقة بالأدلة الجنائية، إذ يمكن استخدامه كقرينة عند حدوث جريمة ووجود تسجيل لصوت له علاقة بها. حيث يقوم الخبير الصوتي باستخدام نظم التعرف على المتحدث إضافة إلى خبرته في تحليل الرسوم الطيفية [١].

٤. الخاتمة

قدم هذا الفصل عرضا عاما لتخصص الصوتيات بفروعه الثلاثة: الصوتيات النطقية والأكوستية والسمعية. التي تشكل الخلفية العلمية للتطبيقات التقنية ذات العلاقة بالكلام. فتطورت تقنيات الاتصالات في نهاية القرن التاسع عشر الميلادي كما وكيفا حتى أصبحت وسائل الاتصال الصوتي بين الناس في كل مكان تقريبا (٩٥٪ من المناطق السكنية على مستوى العالم مغطاة بشبكة اتصال الهاتف الجوال عام ٢٠١٦م (ITU)). وظهرت تطبيقات التواصل الصوتي بين الإنسان والآلة وأصبحت قابلة للاستخدام إما بشكل كامل أو بشكل جزئي (بما في ذلك التعرف على الكلام وتوليده والتعرف على المتحدث) مع بقاء الحاجة قائمة لمزيد من التحسين والتطوير لها.

كثير من الإنجازات العلمية والتقنية في عصرنا الحالي قائمة على تعدد التخصصات multidiscipline بما في ذلك نظم الاتصالات والتعامل الصوتي مع الآلة، ولهذا فإن فرق العمل البحثية والتطويرية في عالمنا العربي تحتاج إلى هذا التكامل في عملها. وهذا يتطلب أن تكون التخصصات الأخرى ذات العلاقة حاضرة في تعليمنا الجامعي حتى

١- الغامدي، منصور بن محمد (١٤٢٧هـ) "البصمة الصوتية": أمد بداية التصويت أنموذجا. المجلة العربية للدراسات الأمنية والتدريب. ٢١. ٤٢: ٨٩-١١٨.

يكون المتخصص في مجال ما ملما بالتخصصات ذات العلاقة بتخصصه. فطالب اللغة العربية مطلع على علاقة الإحصاء والحاسب والهندسة الكهربائية في تخصصه، وكذلك المتخصصين الآخرين في الإحصاء والحاسب والهندسة الكهربائية على دراية بأهمية متخصصي اللغة العربية في تخصصهم. هذا التكامل يخدم التخصصات المترابطة ويرفع من كفاءة مخرجاتها.

وإذا كان العرب من أسس علم الصوتيات قبل ما يقرب من ١٢٠٠ سنة، وهم الآن بعيدون عن مستجدات هذا العلم، فإن جامعاتنا وأقسام اللغة العربية فيها أحوج ما تكون لتطوير مساراتها التعليمية لتواكب التطور التقني والعلمي مما يسهم في توفير طاقات بشرية قادرة على تقديم إضافة في مجال تخصصاتهم على المستوى العالمي وعلى مستوى اللغة العربية التي طورت تقنيات لخدمتها من خارج بيئتها مما أضر بها.

التحليل والمعالجة الآلية لنصوص وأصوات اللغة أصبحت مع كل فرد تقريبا في مجتمعنا الذي لا يستطيع الاستغناء عن تطبيقاتها سواء على الحاسب الشخصي أو الأجهزة الكفية. هذه التقنيات لا زالت في بداياتها، وهناك عمل مستمر لتطويرها نظرا للكم الكبير من المحتوى اللغوي المتاح الآن على الإنترنت، ولحاجة المستخدم لمزيد من الأدوات للبحث فيه والاستفادة منه وإثرائه والحضور المؤثر فيه.

مصطلحات عربية/ إنجليزية

عربي	إنجليزي
ألوفون	Allophone
لثة	Alveolar ridge
شدة الموجة الصوتية	Amplitude
تطبيقات	Applications
إدغام	Assimilation
لسانيات حاسوبية	Computational linguistics
صامت	Consonant
السمات المميزة	Distinctive features
نطاق رنيني	Formant
تردد الموجة الصوتية	Frequency
تردد أساس (تردد الرقيقتين الصوتيتين)	Fundamental frequency
قرافيم	Grapheme
حنك صلب	Hard palate
توافقيات	Harmonics
هيرتز (دورة كاملة لموجة صوتية)	Hertz
تقنية المعلمات	Information technology
وضوح في الكلام	Intelligibility

إنجليزي	عربي
International Phonetic Alphabet	الألفبائية الصوتية الدولية
Language identification	تعرف على لغة
Lingua/tongue	لسان
Linguistic level	مستوى لغوي
Linguistics	لسانيات/ لغويات
Lips	شفتان
Morphology	صرف
Naturalness	طبيعية
Oscilloscope	أوسولوسكوب (جهاز لعرض الموجة الصوتية)
Phone	صوت
Phoneme	فون
Phonetics	علم الصوتيات
Phonetic level	مستوى فونتيكي/ صوتي
Phonograph	فونوقراف (بكم) جهاز تسجيل قديم
Phonological rules	قوانين فونولوجية
Phonology	علم الفونولوجيا
Resonance	رنين

عربي	إنجليزي
تقطيع (الموجة الصوتية)	Segmentation
دلالة	Semantics
صوت	Sound
نظام صوتي	Sound system
تعرف على المتحدث	Speaker verification/identification
كلام (الموجات الصوتية اللغوية)	Speech
تحليل الكلام	Speech analysis
معالجة الكلام	Speech processing
تعرف آلي على الكلام	Speech-to-text/Automatic speech Recognition
جهاز المطياف	Spectrograph
رسم طيفي	Spectrogram
نحو/ قواعد لغة	Syntax/Grammar
توليد آلي للكلام	Text-to-speech/speech synthesis
لثة	Tooth
ترميز (وضع الرمز المقابل للموجة الصوتية)	Transcription
لهاة	uvula
حنك لين	Velum

عربي	إنجليزي
رقيقتان صوتيتان	Vocal folds
جهاز صوتي	Vocal tract
صائت	Vowel

المراجع العربية

◇ **أحمد، أحمد راغب أحمد** (٢٠١٤) قضايا خلافية في ضوء التحليل الصوتي الحاسوبي. المؤتمر الرابع عشر حول هندسة اللغة، ٣-٤، ديسمبر ٢٠١٤.

◇ **الغامدي ، منصور بن محمد** (١٤٢٧هـ، أ) تصميم رموز حاسوبية لتمثيل الفبائية صوتية دولية تعتمد على الحرف العربي. مجلة جامعة الملك عبد العزيز: العلوم الهندسية. ١٦، ٢: ٢٧-٦٤.

◇ **الغامدي ، منصور بن محمد** (١٤٢٧هـ، ب) «البصمة الصوتية»: أمد بداية التصويت أنموذجا. المجلة العربية للدراسات الأمنية والتدريب. ٢١. ٤٢: ٨٩-١١٨.

◇ **الغامدي، منصور بن محمد** (١٤٣٦) الصوتيات العربية والفونولوجيا. مكتبة التوبة، الرياض.

◇ **الغامدي، منصور بن محمد، حسني المحتسب، مصطفى الشافعي** (١٤٢٤هـ) قوانين الفونولوجيا العربية، مجلة جامعة الملك سعود: علوم الحاسب والمعلومات. ١٦: ١-٢٥.

المراجع الأجنبية

◇ **Alghamdi, Mansour, Yahia El Hadj, Mohamed Alkanhal** (2007) A Manual System to Segment and Transcribe Arabic Speech. IEEE International Conference on Signal Processing and Communication (ICSPC07). Dubai, UAE: 24–27 November 2007.

◇ **Almosallam, Ibrahim, Atheer AlKhalifa, Mansour Alghamdi, Mohamed Alkanhal, Ashraf Alkhairy** (2013) SASSC: A standard Arabic single speaker corpus. 8th ISCA Synthesis Workshop, Barcelona, Spain.

◇ **Al-Nassir, A.A.** (1985) Sibawayh the phonologist: A Critical Study of the phonetic and phonological Theory of Sibawayh as presented in His Treatise on Al/Kitab. York, D. phil.

◇ **Alotaibi, Yousef A, Sid-Ahmed Selouani, Mansour M Alghamdi, Ali H Meftah** (2012) Arabic and English speech recognition using cross-language acoustic models. 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA), 2012.

◇ **Chang, Y.-Y.** (2011) Evaluation of TTS systems in intelligibility and comprehension tasks. In Proceedings of the 23rd Conference on Computational Linguistics and Speech Processing.

◇ **Chowdhury, Gobinda G.** (2003) Annual Review of Information Science and Technology, v37 p51-89.

◇ **Hanumanthappa, M, Rashmi S, Jyothi N IJISET** (2014) Impact of Phonetics in Natural Language Processing: A Literature Survey. International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 3.

◇ **Islam MA, Jassim WA, Cheok NS, Zilany MSA** (2016) A Robust Speaker Identification System Using the Responses from a Model of the Auditory Periphery. PLoS ONE 11(7): e0158520. doi:10.1371/journal.pone.0158520

◇ **Ismail1, Hassan. N. A., M. Hesham Farouk El-Sayed, M. H. Eid, Wael A. Sultan** (2015) Statistical Models for Arabic Phonemes Recognition. The 04th International Conference for Statistics, Computer Science and its Applications.

◇ **Juang, B.H., and Lawrence R. Rabiner** (2005) Automatic Speech Recognition – A Brief History of the Technology Development. Encyclopedia of Language and Linguisics, Elsevier.

◇ **Jonathan Owens** (2013) The Oxford Handbook of Arabic Linguistics. Oxford University Press.

◇ **Nahar, Khalid M. O., Mohammed Abu ShquierWasfi G. Al-KhatibHusni Al-MuhtasebMoustafa Elshafei** (2016) Arabic phonemes recognition using hybrid LVQ/HMM model for continuous speech recognition. International Journal of Speech Technology, 19: 03.

◇ **Nakai, Satsuki, David Beavan, Eleanor Lawson, Grégory Leplâtre, James M Scobbie & Jane Stuart-Smith** (2016) Viewing speech in action: speech articulation videos in the public domain that demonstrate the sounds of the International Phonetic Alphabet (IPA). Innovation in Language Learning and Teaching.

◇ **Ohala, J. J.** (1991) The integration of phonetics and phonology. International Congress of Phonetic Sciences, Aix-en-Provence, 19-24 Aug 1991. Vol. 1, pp. 1-16.

◇ **Robjohns, Hugh** (2001). "A Brief History of Microphones". Microphone Data Book. Archived from the original on 2010-11-25.

◇ **Ruslan Mitkov** (2009) The Oxford Handbook of Computational Linguistics. Oxford University Press.

◇ **Singh, Rita, Joseph Keshet, Deniz Gencaga, and Bhiksha Raj** (2016) The relationship of voice onset time and Voice Offset Time to physical age .IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016.

◇ **Wu, Zhizheng, Oliver Watts, Simon King** (2016) Merlin: An Open Source Neural Network Speech Synthesis System. 9th ISCA Speech Synthesis Workshop.

◇ Auditory Neuroscience: https://auditoryneuroscience.com/topics/formants-and-harmonics-spoken-vowels

◇ HTK: http://htk.eng.cam.ac.uk

◇ ITU:

http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2016.pdf

◇ Universiteit Leiden: http://fonetiek.ullet.net/pacilly/Pras/vk-6-sg-6061-b.jpg

◇ Wikimedia:

https://commons.wikimedia.org/wiki/File:RAV4_Tach_750rpm.JPG

◇ Wikipedia: https://en.wikipedia.org/wiki/Siri

◇ ◇ ◇

الفصل الثاني

التحليل الصرفي

د. عبدالعزيز بن عبدالله المهيوبي[1]

ملخص البحث

يُعالج هذا المبحث قضيَّة مهمة من قضايا معالجة الصرف العربي حاسوبياً، وهي (التحليل الصرفي الآلي للغة العربية)، وقد تناولها الباحث في البداية بعرض موجز لخصائص الصرف العربي، وانتقل منها إلى الحديث عن مفهوم التحليل الصرفي الآلي، وقواعد المعطيات المصاحبة للمحلل الصرفي. كما اقترح الباحث مجموعة من الأسس المهمة لبناء محلل صرفي دقيق للغة العربية. كما قدّم نظرة تاريخية للتحليل الصرفي الآلي، وعرض لمجموعة من أهم المحللات الصرفية العربية. وتطرق لأهمية التطبيقات الحاسوبية للتحليل الصرفي.

١ – أستاذ اللغويات الحاسوبية المساعد في جامعة الإمام محمد بن سعود الإسلامية. درس الدكتوراه في قسم علم اللغة التطبيقي بمعهد تعليم اللغة العربية. له عدة أبحاث منشورة حول معالجة اللغة العربية آلياً. شارك في العديد من المشروعات البحثية، كمشروع تعليم اللغة العربية عن بعد، ومشروع إعداد اختبار كفاية لغوية في اللغة العربية للناطقين بغيرها وهما مشروعان لغويان تفاعليان بين عمادة التعلم الإلكتروني والتعليم عن بعد، ومعهد تعليم اللغة العربية بجامعة الإمام محمد بن سعود الإسلامية، عضو هيئة تحرير مجلة اللسانيات العربية وأمينها، صاحب موقع اللسان العربي على الإنترنت http:// www.lisan2.com. مهتم بتقنيات التعليم والتصميم التعليمي والتقني.(s99scom@hotmail.com)

وانتقل الباحث بعد ذلك إلى الحديث عن مجموعة من الضوابط والمحددات التي تساعد في بناء المحللات الصرفية، مقسِّماً إيّاها إلى ضوابط ومحددات شكليَّة ودلالية. ولم يغفل الباحث الحديث عن المشكلات التي تواجه بناء محلل صرفي دقيق لكلمات اللغة العربية ونصوصها، وطرق عرض نتائجها، وكيفية توصيف القواعد الصرفية لبناء المحلل الصرفي الآلي. وأشار في عجالة إلى أسباب قصور المحللات الإنجليزية عن استيعاب خصائص اللغة العربية، وتحدث بالتفصيل عن خطوات بناء المحلل الصرفي الآلي، ومتطلبات بنائه. وانتهى الباحث إلى أنَّ النجاح في تطوير تطبيقات حاسوبية للتحليل الصرفي يتوقف على مدى وضوح النظرية اللغوية التي يتبناها مطورو المحللات الآلية، ومدى قدرتها على تحليل الكلمات والنصوص تحليلاً صحيحاً.

الفصل الثاني: التحليل الصرفي

منذ ظهور الجيل الأول للحواسيب في عام ١٩٤٤م، وصلة الحواسيب باللغات الطبيعية تزيد وتتطور "وذلك لسبب أساسي وبسيط، وهو كون اللغة تجسيد لما هو جوهري في الإنسان، أي نشاطه الذهني بكل تجلياته، في الوقت نفسه الذي تتجه فيه الحواسيب نحو محاكاة بعض وظائف الإنسان، وقدراته الذهنية"[١]. وقد أولت الأمم المتقدمة أهمية كبيرة في عصر الرقمنة لمعالجة لغاتها الطبيعية بواسطة الحاسوب، ويُحتِّم هذا الأمر علينا دراسة لغتنا العربية محاولين توصيف قواعدها، وميكنتها بالحاسوب، مستفيدين من خصائصها في تطوير برمجيات، وبناء قواعد معطيات تساعد في معالجتها آلياً، «فمنذ الأربعينيات والمحاولات مستمرة لتحوير قواعد اللغات الطبيعية من الشكل الوصفي الأدبي إلى الشكل العلمي الدقيق، والذي يمكن برمجته حسب الإمكانيات التي يوفرها الحاسوب...، وتتوفر الآن لغات برمجة عالية المستوى تتسم بها يُطلق عليه الذكاء الاصطناعي، الذي يجعل الحاسوب يستقبل، ويحلل، ويُنفذ ما يُعطى إليه»[٢].

١- نبيل علي، اللغة العربية والحاسوب (دراسة بحثية)، تعريب، د.ط، ١٩٨٨م، ص١١٤.

٢- يُنظر: عبده ذياب العجيلي، الحاسوب واللغة العربية، منشورات جامعة اليرموك- عمادة البحث العلمي والدراسات العليا، الأردن، د ط، ١٩٩٦م، ص١٤.

وتمرُّ اللغات الطبيعية بمراحل لفهم الكلام، أهمها: التحليل الصرفي - وهو موضوع اهتمامنا هنا - والتحليل المعجمي، والتحليل الدلالي[١]، والتحليل النحوي[٢]. والعربية لغة قابلة للمعالجة الحاسوبية» إذ تملك نظاماً خاصاً يجعلها أكثر قابلية لأن تُمَثَّل حاسوبياً، وتوفر اللغة العربية مجالات كثيرة لتناولها حاسوبياً على اختلاف مستوياتها، ولاسيما الصرفية؛ لأنها لغة ذات نظام دقيق تركيبياً، ودلالياً، ومعجمياً»[٣]. لقد أثبتت اللغة العربية أنها من أكثر اللغات قابلية لاستخدام الحاسوب في معالجتها آلياً؛ «لأنها تجمع بين كثير من الخصائص اللغوية المشتركة مع اللغات الأخرى، فأبجديتها -مثلاً- ليست فونيمية صرفة كالإسبانية والفنلندية، حيث يقابل كلُّ حرف صوتاً واحداً، كما أنها ليست مقطعية كاليابانية، حيث رموز الأبجدية عبارة عن مقاطع يتكون كل منها من صامت (حرف) يتبعه صائت (حركة) مثل: «ما، كي، فو» فالأبجدية العربية رغم كونها فونيمية أساساً، فإنها تتضمن حروفاً ذات طبيعة مقطعية، مثل: «لا، لأ، ؤ»[٤].

وتتعرض لغتنا العربية لهجوم كبير، ورَمْي بالقصور والعجز عن مواكبة التطور، وذلك بهدف تدمير اللغة التي يتعبد بها أكثر من مليار مسلم، ناهيك عن كونها الدعامة الأساسية، والمعبر الحقيقي للقوة التي تستند إليها الحضارة العربية، والتراث الإسلامي. ولعل آخر هجوم عليها هو عدم قابليتها للمعالجة الآلية باستخدام الحاسوب.

«وعلم الصرف من أهم العلوم العربية قديماً وحديثاً؛ فلا يمكن لنحوي، أو لغوي، أو معلم، أو طالب الاستغناء عنه؛ لأنه أساس العربية، وميزانها، به تتولد الكلمات[٥]، وبه يتم الاشتقاق، كاسم الفاعل، واسم المفعول، والصفة المشبهة، وصيغة المبالغة

١- يستخلص المحلل الدلالي معاني الكلمات استناداً إلى سياقها، ويحدد معاني الجمل استناداً إلى ما يسبقها، وما يلحقها من جمل.

٢- يحدد المحلل النحوي صيغة الفعل في الماضي والمضارع والأمر، وكونه مؤكداً، أو غير مؤكد، ومعلوماً أو مجهولاً، ومبنياً أو معرباً، وعلامة كل منهما، ظاهرة أو مقدرة، حرفاً كانت أو حركة، ويبيِّن ضمائر الرفع أو النصب التي أسند إليها الفعل، كما يبيّن المحلل النحوي حالة الاسم من حيث الإعراب والبناء، وعلامة كلّ منهما، أمّا الحرف فيبيّن المحلل سابقته، ولاحقته، ووظيفته الإعرابية، وعلامة بنائه.

٣- مازن الوعر، دراسات لسانية تطبيقية، دار طلاس، دمشق، ط١، ١٩٨٩م، ص٣٧٧.

٤- سعيد أحمد بيومي، أم اللغات -دراسة في خصائص اللغة العربية، والنهوض بها-، ط١، ٢٠٠٢م، ص١٠٥.

٥- أنواع الكلمات في اللغة العربية: جذور بدون إضافات، وجذور تضم سوابق فقط، وجذور تضم لواحق فقط، وجذور تضم سوابق ولواحق، وجذور تضم أواسط فقط، وجذور تضم أواسط وسوابق فقط، وجذور تضم أواسط ولواحق فقط، وجذور تضم أواسط وسوابق ولواحق.

وغيرها، وبه يُعرف الصحيح من المعتل، والمجرد من المزيد، والأوزان المختلفة، وبه تُعرف الأسماء تعريفاً وتنكيراً، وجنساً وعدّاً، ناهيك عن أنَّه يقي اللسان من الوقوع في الخطأ، ويرشده إلى الصواب، ويصحِّح القلم من الزّلات، وتتكئ عليه الحقول اللغوية المختلفة، الصوتية، والنحوية، والتركيبية، والمعجمية، والدلاليّة[1]».

«والصرف هو رابطة العقد لعناصر المنظومة اللغوية، فهو ركيزة الفونولوجي، ومدخل النحو، وأساس تنظيم المعجم، وفوق هذا كله فهو خط المواجهة الساخن لالتقاء مباني اللغة ومعانيها»[2]. ويُعد الصرف في اللغة العربية مصدر التوسع اللغوي بما يوفره من وسائل عديدة لتكوين كلمات جديدة، وإعادة تحليل تلك الكلمات، «كما يُعد الصرف العربي وضعاً مثاليّاً لإبراز ثنائية التحليل والتركيب[3]، وذلك نظراً لكون كلٍّ منهما صورة منعكسة من الآخر بصورة لا تتوافر في الفروع اللغوية الأخرى»[4].

وتتسم اللغة العربية باطّراد نظام صرفها، وظهور أثره داخل منظومتها اللغوية «لذا فهو مدخل أساسي لوصف النظام الشامل للغة العربية، وتفسير الكثير من ظواهرها، وتحديد أسلوب معالجتها آليّاً»[5]. ويُعد التحليل الصرفي إحدى مراحل معالجة اللغات الطبيعية آليّاً، ويدخل في الكثير من التطبيقات اللغوية، مثل: ميكنة المعاجم، وضغط النصوص، وتشكيلها، وتشفيرها، وتحليلها، وتمييز الكلام، وتوليده، وتصحيح الأخطاء الإملائية، والترجمة الآلية.

وأكاد أجزم أنَّ لمعالجة الصرف آليّاً دور حيوي في جميع الأمور المتعلقة بتناول اللغة العربية حاسوبيّاً ومعلوماتياً؛ حيث تُعد ميكنة العمليات الصرفية بالنسبة للغة العربية مدخلاً أساسياً، وقاسماً مشتركا لمعظم نظمها الآليّة، حيث يزعم الدكتور «نبيل علي» «أنَّ مدى نجاحنا في تعريب نظم المعلومات والمعارف، يتوقف بالدرجة الأولى على

١- محمود مصطفى عيسى خليل، إسناد الأفعال إلى الضمائر في ضوء اللسانيات الحاسوبي- ماجستير، كلية الدراسات العليا، الأردن، ٢٠١١م، ص٤٥.

٢- نبيل علي، اللغة العربية والحاسوب (دراسة بحثية)، تعريب، د.ط، ١٩٨٨م ، ص٢٤٧.

٣- وهي الثنائية التي يوصف من خلالها كثير من الظواهر اللغوية في الوقت نفسه الذي تُعدُّ فيه أحد المفاهيم الأساسية في تصميم نظم المعالجة الآلية للغات.

٤- يُنظر: نبيل علي، اللغة العربية والحاسوب (دراسة بحثية)، تعريب، د.ط، ١٩٨٨م ، ص ٢٤٧-٢٤٨.

٥- المرجع السابق ، ص٢٤٨.

ما نستطيع أن نحققه على جبهة الصرف»[١]، ويقصد الدكتور «نبيل علي» هنا الصرف بمعناه الواسع: مبناه ومعناه، تصريفه وتركيبه، تحليله وتوليده، اطِّراده وشذوذه.

وحتى نتمكن من تحليل اللغة العربية نحتاج إلى معرفة مفرداتها، وطريقة تركيب تلك المفردات في سياقات للحصول على جمل مفيدة. كذلك نحتاج إلى معرفة معاني تلك المفردات، وطرق استعمالها في الكلام.

كما أنَّ الصرف هو المسؤول عن بنية مفردات اللغة، تحليلاً وتوليداً. ويتلقى الصرف مدخلاته في اللغة العربية من ثلاثة مصادر، حيث تتسم المنظومة اللغوية بالتماسك الشديد بين عناصرها، وهذه المصادر هي: المعجم: حيث يُغذي الصرف بجذور المفردات، أو جذوعها، والدلالة: حيث تحدد المعنى الصرفي المراد صياغة الكلمة في قالبه، والنحو: حيث يعيّن الوظيفة النحوية للمفردة داخل الجملة، وحالتها الإعرابية»[٢].

١. خصائص الصرف العربي[٣]

تتسم اللغة العربية بخاصية الاشتقاق الصرفي المبني على أنماط الصيغ، إذ إنها تتميز بالاطَّراد الصرفي المنتظم الذي أدى بالبعض إلى وصفها بالجبرية (نسبة إلى علم الجبر) بدرجة تقترب من حد الاصطناع. كما تتميز اللغة العربية بالتعالق الشديد بين مستوياتها، حيث يتعالق المستوى الصرفي مع المستوى الصوتي، فيعتمد الصرف اعتماداً كبيراً على نتائج علم الأصوات عند الحديث -مثلاً- عن الإعلال والإبدال.» كما أن النحو لا يتخذ لمعانيه مباني من أي نوع إلا ما يقدمه له الصرف من المباني، وهذا هو السبب الذي جعل النحاة يجدون في أغلب الأحيان أنه من الصعب أن يفصلوا بين الصرف والنحو، فيعالجون كلاً منهما علاجاً منفصلاً، ومن هنا جاءت متون القواعد مشتملة على مزيج من هذا وذاك، يصعب معه إعطاء ما للنحو للنحو، وما للصرف للصرف»[٤]. وقد ارتبطت عملية الكشف على المعاجم بعملية التحليل الصرفي، علاوة على ذلك فالتماسك المعجمي ممثلاً في الاشتقاق، وكذلك في العلاقات الدلالية بين

١- المرجع السابق ، ص٢٩٧.

٢- المرجع السابق ١٩٨٨م ، ص٢٩.

٣- عبدالعزيز بن عبدالله المهيوبي، بناء خوارزمية حاسوبية لتوليد الأفعال في اللغة العربية وتصريفها - دكتوراه، معهد تعليم اللغة العربية - جامعة الإمام محمد بن سعود الإسلامية، ١٤٣٦هـ، ص١١٨.

٤- تمام حسّان، اللغة العربية معناها ومبناها، دار الثقافة، المغرب، ١٩٩٤م، ص١٧٨.

المفردات المشتركة في الصيغة الصرفية الواحدة، هو نتيجة طبيعية لشدة التماسك بين الصرف والمعجم.

وسنركِّز عند حديثنا عن خصائص الصرف العربي على تلك النواحي ذات الصلة بمعالجته آلياً، حيث تعد معالجة الصرف العربي آلياً مطلباً أساسياً لميكنة عمليات تحليل النصوص المكتوبة والمنطوقة، وفهمها وتوليدها، علاوة على أنّه أساس لا غنى عنه لميكنة المعاجم واسترجاع المعلومات وتحليل مضمون النصوص. حيث يتميز الصرف العربي بعدة خصائص من أهمها:

١- وضوح مسار عملية الاشتقاق (الانتقال من الجذور إلى المشتقات الفعلية).

٢- اطِّراد التصريف في العربية، باستثناء حالات نادرة.

٣- ميل الصرف العربي لتركيب الكلمات بالإضافة، وكرهه لتكوين الكلمات من خلال المزج والاختصار.

٤- انتظام بنية الكلمة العربية لثبوت رتبة عناصرها (الصرف-نحوية).

٥- شدة التداخل بين الصرف، والفونولوجي من حيث تعدد قواعد الإبدال والإعلال، وعمليات التغيير (الصرف-صوتية) الأخرى.

٦- قلة عدد جذور الأفعال وكثرة عدد فروعها.

٧- أن الاشتقاق في العربية مبني على الأنماط الصرفية [1]، حيث تتعدد هذه الأنماط مستخدمة عدداً قليلاً من حروف الزيادة.

٨- محورية مفهوم الجذر في العربية كعنصر ربط معجمي ودلالي.

٢. الحاسوب ومحاكاة تفكير الإنسان

سعى علماء اللسانيات الحاسوبية إلى بناء تطبيقات وأدوات للتحليل الصرفي الحاسوبي؛ بهدف محاكاة التفكير الإنساني في تحليل كلمات ونصوص اللغات الطبيعية من النواحي الإدراكية والنفسية. ولكن هل تمكَّنوا من ذلك؟ الجواب: لا، لأنَّ علماء اللسانيات الحاسوبية لم يتمكنوا من بناء محلل صرفي متكامل يحاكي تفكير الإنسان،

١- النمط الصرفي:عبارة عن قالب يشمل الحركات وحروف الزيادة ومواضع حروف الجذر بتسلسل ورودها داخل القالب.

على الرغم من كل المحاولات الجادة التي تُبذل لتحقيق هذا الهدف، محاولين «استكناه العمليات اللاإرادية التي تحدث في العقل البشري التي يمكن من خلالها إعطاء الحاسوب القدرة على فهم اللغة الطبيعية، وتحليلها، وإعادة إنتاجها، وكيفية تشكيلها في العقل البشري» [١].

٣. التحليل الصرفي

«يُقصد بالتحليل الصرفي الآلي للكلمة في اللغة العربية «ربط كلمات النص بالعناصر الصرفية الأولية التي تدخل في تكوينها، وكذلك بالقيم النحوية دون اعتبار موقعها»[٢]. فيتم في التحليل الانتقال من الكلمة إلى جذرها الأصلي؛ أي أنَّ الحاسوب يعالج الكلمات العربية المشكولة جزئياً، أو كلياً، أو غير المشكولة، فيصف ما يطرأ عليها من تغيير؛ زيادة، أو نقصاناً، أو إعلالاً، أو إبدالاً، أو إدغاماً، أو قلباً، حيث «يحدد نوعها، وميزانها الصرفي، وسابقتها (أو سوابقها)[٣]، ولاحقتها (أو لواحقها)[٤]، وحالتها الإعرابية، ودلالتها، ... فإذا احتوت الكلمة المراد تحليلها على حروف غير مشكولة، وضع الحاسوب الحركات الممكنة لها اعتماداً على بيانات مخزنة. ومن المعلوم أن خلوَّ الكلمة من الشكل يجعلها متعددة الأشكال، ومن ثمَّ المعاني، مادامت مستقلة عن سياق النص.»[٥]

«فكلمة (وجد) مثلاً يمكن أن تكون لها الإمكانيات التالية:

وَجَدَ، وَجَّدَ، وُجِدَ، وُجِّدَ ... = أفعال / وَجْدٌ = اسم / وَ+جَدَّ، وَ+جُدَّ ... = حرف عطف+أفعال / وَ+جَدٌّ = حرف عطف+اسم.

ومع ذلك فالكلمة المشكولة إذا عوجلت مستقلة عن سياق النص، فلا يمنع شَكْلُها

١- نهاد الموسى، العربية نحو توصيف جديد في ضوء اللسانيات الحاسوبية، المؤسسة العربية للدراسات والنشر، بيروت، ط١، ٢٠٠١م، ص٥٧.

٢- يحيى هلال، التحليل الصرفي للعربية، وقائع مختارة من ندوة استخدام اللغة العربية في الحاسب الآلي في الكويت، عمّان، دار الرازي، ص٢٦٦.

٣- السوابق مجموعة من الحروف، والأدوات التي تسبق الكلمة، وتؤدي إلى تغيير معناها، أو وظيفتها النحوية.

٤- اللواحق مجموعة من الحروف التي تُضاف إلى آخر الكلمة، فتُغير معناها، أو وظيفتها النحوية.

٥- يُنظر: مروان البواب، ومحمد الطيّان، أسلوب معالجة اللغة العربية في المعلوماتية (الكلمة - الجملة)، استخدام اللغة العربية في المعلوماتية.

من إمكانية اشتراكها في الاسمية والفعلية، أو الفعلية والحرفية، فمن أمثلة الحالة الأولى كلمة (يَزِيدُ) فهي اسم في نحو قولنا: خَرَجَ يزيدُ من الغرفة. وفعلٌ في نحو قولنا: يزيدُ الله في خلقه ما يشاء. ومن أمثلة الحالة الثانية كلمة (أنَّ).

وهذا يعني أن على الحاسوب أن يعالج الكلمة عند تحليلها على أنها فعلٌ واسمٌ وحرف، وأن يعطي جميع الإمكانيات المحتملة لها، مع مراعاة الحالات التي تحدد نوعها، فالكلمة المنونة -مثلاً- لا تكون إلا اسماً. وبعد ذلك يقوم الحاسوب باختيار الإمكانية المناسبة التي تتوافق مع سياق النص»[١].

٤. المحلل الصرفي الآلي

هو تطبيق حاسوبي يقوم باستخلاص العناصر الأولية لبنية الكلمة في اللغة العربية، ويُحدد سماتها الصرفية، والصرف صوتية، والصرف نحوية، فيقوم المحلل الآلي بالكشف عن جذر الكلمة، ووزنها الصرفي، ويبيِّن ما يطرأ عليها من تغيير بالزيادة أو النقصان، والإعلال، والإبدال، والإدغام، والقلب، ويوضح ما يلحقها من سوابق، ولواحق، وزوائد، بالإضافة إلى تقسيم الكلمة إلى اسم، أو فعل، أو حرف، وتقسيم الاسم إلى جامد، أو مشتق، ومذكر، أو مؤنث، ومفرد أو مثنى أو جمع ...إلخ.

ويضم المحلل الصرفي مجموعة من قواعد المعطيات: هي قواعد معطيات معجمية لأوزان الكلمات العربية المستعملة، وقواعد معطيات لأسماء الأعلام، وقواعد معطيات للأخطاء الإملائية، والنحوية الشائعة في نصوص اللغة العربية.

٥. توأمة النحو والصرف في المعالجة الآلية

التداخل الكبير بين المستويين اللغويين الصرفي والنحوي في اللغة العربية «أوجب معالجتهما آلياً بأسلوب متداخل، فهناك تداخل مستمر بين الصرف والنحو ينشأ في بداية عملية التحليل الصرفي الآلي للكلمات والجمل، ويستمر حتى نهايتها»[٢]. فالمستويات اللغوية في اللغة العربية رغم تباينها فهي «متداخلة متكاملة دون تفاضل أو تمايز، فهي

١- عبدالعزيز بن عبدالله المهيوبي، بناء خوارزمية حاسوبية لتوليد الأفعال في اللغة العربية وتصريفها - دكتوراه، معهد تعليم اللغة العربية - جامعة الإمام محمد بن سعود الإسلامية، ١٤٣٦هـ، ص١٢٦ - ١٢٧.

٢- هدى آل طه، النظام الصرفي للعربية في ضوء اللسانيات الحاسوبية "مثل من جمع التكسير"، رسالة دكتوراه، الجامعة الأردنية، ٢٠٠٥م، ص٢١.

كلٌّ واحد تتآزر في بناء اللغة؛ فاللغة كالجسد الواحد، تربطه وحدة عضوية، وتصل بين أجزائه شرايين، وأعصاب قد تكون متناهية الدقة، إلا أنها تجعل سلامة عضو تعتمد على سلامة عضو آخر، بل حياته أيضاً»[(١)]، فالصرف يعتمد على الأصوات من جهة، وعلى النحو من جهة أخرى، فالعلاقة بين المستويين الصرفي والنحوي وطيدة، فهما توأمان يصعب الفصل بينهما.

٦. أهمية التحليل الصرفي

تتجلى في المحلل الصرفي أهم خصائص اللغة العربية في مجال المعالجة الحاسوبية، فيُتاح عن طريق المحلل الصرفي توليد جميع الكلمات التي يمكننا اشتقاقها من جذر معين، كما نستطيع من خلاله ردَّ أيَّ كلمة مشتقة إلى جذرها، أو أصلها الذي تعود إليه. كما يستطيع المحلل بعد الكشف عن جذر الكلمة توليد الأسماء المشتقة من الفعل المجرد، أو المزيد، ويولِّد مزيدات الفعل الثلاثي بحرف وبحرفين وبثلاثة أحرف، وكذلك مزيدات الفعل الرباعي بحرف وبحرفين، ويكتشف ما يصيب الكلمة من حالات الإعلال، أو الإبدال، أو الهمز، أو التضعيف.

تُعدُّ تطبيقات التحليل الصرفي لكلمات اللغة العربية بمثابة الأساس والقاعدة للتطبيقات الحاسوبية اللغوية الأخرى، إذ تستفيد منها بشكل، أو آخر، ولكنَّها تصبح أساسية بالنسبة لتطبيقات البحث والفهرسة، فهي تطبيق مباشر لها، حيث يقف المحلل الصرفي في مكان الصدارة بوصفه التطبيق الفاعل والسريع للمساعدة في الوصول إلى الكلمات المطلوبة عن طريق إعادة الكلمة المشتقة إلى جذرها، والتعرّف على الصور الصرفية لها. كما يُستخدم المحلل الصرفي في الترجمة الآلية، واسترجاع البيانات، "فيتولى المحلل ربط المفردات المختلفة للصيغ، مثل (كتب، الكتب، يكتبون، كاتبون، كتبتُ ...) التي يمكن استرجاعها تحت الجذر (ك ت ب) بالإضافة إلى إمكانية استرجاع الكلمات المختلفة حسب صياغاتها المتفاوتة، مع ما يتصل بها من سوابق أو لواحق"[(٢)].

كما تبرز أهمية المحلل الصرفي عند التعامل مع النصوص العربية الكبيرة، مثل القرآن الكريم، وموسوعات الحديث النبوي على الحاسوب، فيكفي على سبيل المثال

١- نبيل علي، اللغة العربية والحاسوب (دراسة بحثية)، تعريب، د.ط، ١٩٨٨م ، ص٤٠٣.

٢- علي السليمان الصوينع، استرجاع المعلومات في اللغة العربية، مطبوعات مكتبة الملك فهد الوطنية، السلسلة الثانية، الرياض، ١٩٩٤م، ص١٤٠.

أن تستخدم للبحث في القرآن الكريم جذراً، مثل «س ل م» فيستدعي المحلل جميع الآيات القرآنية التي وردت بها مشتقات هذا الجذر، مثل: (أسلم، سلام، سالمون، سليم، مسلمون، الإسلام ..إلخ)

ويُعد المحلل الصرفي إحدى الدعائم الأساسية التي يقوم عليها مشروع خدمة السنة النبوية، فحاجتنا للبحث - على مستوى الجذر- في الأحاديث التي تضم عدداً كبيراً من الألفاظ تتضاعف مع كثرة البحث، وتعدد أهدافه»[١]، كما يمكن للمحلل الصرفي دعم التشكيل الآلي للكلمات الخالية من التشكيل، حيث يعطي مجموعة من الخيارات لتشكيل الكلمة داخل النص. ويُساعد في التدقيق الإملائي للنصوص العربية؛ حيث يكتشف الخطأ الإملائي، ويقترح البدائل الصحيحة المحتملة، فالمحلل الصرفي عندما لا يكتشف الساق السليم للكلمة، والصيغة الصرفية التي يجوز انطباقها عليه، فإنه يَعدُّ الكلمة خاطئة، ثمَّ يقدم عدة احتمالات لتصويب الكلمة الخاطئة، عن طريق توليد احتمالات صحيحة، بحيث تكون أقرب ما تكون للصيغ الصرفية السليمة.

كما تدعم تطبيقات التحليل الصرفي محركات البحث في الإنترنت؛ حيث يمكننا البحث عن كلمة أو جملة أو مجموعة كلمات بحث مطابق، أو باللواصق، أو على مستوى الجذر، فإذا أردنا البحث عن كلمة (رأى) وكل ما يرتبط بها من كلمات داخل صفحات الشبكة العالمية، فسنحصل - بمساعدة تطبيقات التحليل الصرفي - على قائمة طويلة من الكلمات التي لا تشترك في بداياتها، أو نهاياتها، ولكنها تشترك في الجذر، مثل (نرى، يريكم، أرنا، يرون، تَرَ، أرني، رأيتموه، أراكم، رأيت، أراك، ليريه، فترى، يروا، أرأيتكم، ليريها، ترونهم، تراني، سأريكم، رأوا..إلخ.

٧. الهدف من بناء محللات صرفية آلية للغة العربية

إنَّ الهدف من بناء المحللات الصرفية الآلية هو بناء أداة لغوية تُمَكِّنُ الحاسوبَ من مشابهة الإنسان في كفايته، وأدائه اللغويين؛ «ليكون قادراً على تحليل نصوص اللغة العربية، وكلماتها، فيكتشف الأخطاء الإملائية عن طريق معرفة النظام الكتابي للغة العربية، ويحلل الصيغ الصرفية، ويتعرفها في سياق الكلام»[٢].

١- محمود عوض المراكبي، تطويع تقنية المعلومات لخدمة العلوم الشرعية، السجل العلمي لندوة استخدام اللغة العربية في تقنية المعلومات، الرياض، ١٩٩٢م، ص١٣٠.

٢- مسفر محماس الدوسري، برمجة الاسم المنسوب بياء النسب في العربية حاسوبياً - رسالة دكتوراه، جامعة اليرموك - كلية الآداب، ٢٠١٠م، ص٦.

٨. عرض نتائج التحليل

تتفاوت المحللات الصرفية في طريقة عرض نتائج تحليل الكلمات والنصوص، وترتيبها، وذلك تبعاً لقوة المحلل، والمنهج المتَّبع في الترتيب، حيث نجد أن بعض المحللات الآلية تكتفي بذكر نوع الكلمة وزوائدها، في مقابل ذلك اهتمَّ بعض مطوري المحللات بذكر سابقة الفعل، ولاحقته، وصيغته الصرفية، وتجرده، أو زيادته، ووزنه، وجذره، وإسناده، وبنائه للمعلوم، أو المجهول، وعلامات البناء، والإعراب، وضبطه بالشكل التام، وسابقة الحرف، ولاحقته، وعلامات بنائه، وسابقة الاسم، ولاحقته، ووزنه، وجذره، ونوعه من جهة التصرف، وعدمه، ومن جهة التذكير، والتأنيث، والنسبة، والتصغير، وعلامات بنائه، وإعرابه، وضبطه بالشكل التام.

٩. خطوات عمل المحلل الصرفي الآلي

عند تحليل الأفعال –مثلاً– فإن الحاسوب سيفترض أنَّ الكلمة المطلوب تحليلها هي فِعْلٌ، «فيقوم بتحديد سوابقها ولواحقها، ويحدد كذلك صيغة الفعل، وبناءه للمعلوم أو المجهول، وهل هو مجرَّد أم مزيد؟ ويحدد وزنه، وأصله المشتق منه، ويُبيِّن حالته الإعرابية، والضمير المسند إليه. أي أنه يعطي وصفاً كاملاً عن حالة الفعل الصرفية، والنحوية، والدلالية مستقلة عن سياق النص. وتمرُّ عملية تحليل الفعل بمراحل كثيرة»[١]، فعند تحليل كلمة «وحزن» يبدأ المحلل بالخطوات التالية:

الخطوة الأولى: وهي خطوة تشذيب[٢] الكلمة المدخلة، حيث ينزع المحلل الزوائد التي لحقت بساق الكلمة، ليكتشف أنها تحتمل أكثر من تحليل، فينزع في الاحتمال الأول (الواو) كحرف عطف ليكون ساق الكلمة هو (حزن)، أمّا في الاحتمال الثاني فيقتطع الحاسوب أحد حروف الفعل الأصلية (النون) ظناً منه أنها لاحقة زائدة لجمع المؤنث، ويُبقي على السابقة (الواو)، ولكنَّه يقوم بإعادتها بعد أن يفشل في تحليل ما تبقى من الكلمة (وحز) فيُعيد السابقة المقتطعة إليها، ثم يحللها مرة أخرى، والاحتمال الثالث هو

١– عبدالعزيز بن عبدالله المهيوبي، بناء خوارزمية حاسوبية لتوليد الأفعال في اللغة العربية وتصريفها – دكتوراه، معهد تعليم اللغة العربية – جامعة الإمام محمد بن سعود الإسلامية، ١٤٣٦هـ، ص١٢٧.

٢– التشذيب هو عملية إزالة كلٍّ من بوادئ الكلمة ولواحقها لإنتاج الجذر او الجذع. وهي طريقة رياضية تجمع كل الكلمات التي تتقاسم الأصل نفسه، وتملك بعض العلاقات الدلالية، حيث تعمل طريقة التشذيب المبنية على الجذع على إزالة السوابق واللواحق المتصلة بالكلمة، في حين تحوِّل الطريقة المبنية على أساس الجذر الأصول إلى جذور.

نزع السابقة (الواو) واللاحقة (النون)، ليكون ساق الكلمة (حز)، وهي ساق سليمة عند تضعيف الحرف الثاني، أمّا الاحتمال الرابع فهو نزع السابقة (الواو) واللاحقة (النون) ليكون ساق الكلمة (حز) وهي ساق سليمة عند حذف حرف العلة، وأصلها (حوز).

الخطوة الثانية: يكتشف المحلل العمليات الصرف صوتية التي تمت على ساق الكلمة، وهي في الاحتمال الأول (حَزَنَّ) بنون النسوة، وفي الاحتمال الثالث (حُزَّنَّ، حُزِّنَّ، حُزُّنَّ) بنون التوكيد الثقيلة، وفي الاحتمال الرابع (حُزْنَ) بنون النسوة، أمّا في الاحتمال الثاني فلا يجد المحلل أية أدلة على حدوث تغيرات صرف صوتية حدثتْ على ساق الكلمة.

الخطوة الثالثة: يقوم الحاسوب بعد ذلك بمقابلة ما تبقى من الفعل مع جداول الأفعال النموذجية المخزنة في الذاكرة. فيبحث عن الأفعال المساوية له في عدد الحروف، ليحصل على جميع الأفعال المفترضة الموافقة لهذا الفعل. وتحتوي جداول الأفعال النموذجية على جميع المعلومات المتعلقة بجذر الفعل، الباب الذي يتصرف منه، ووزنه، ومساره الاشتقاقي، حيث يحدد المحلل جذر الساق، وصيغته الصرفية للاحتمال الأول بأنها (ح ز ن) و (فعِل) أو (فعُل) أو (فَعَّل) أمّا الاحتمال الثالث فجذر الساق هو (ح ز ز) وصيغته الصرفية (فعل)، بالنسبة للاحتمال الرابع فيكون الجذر (ح و ز) وصيغته (فعل)، أما الاحتمال الثاني فلا تسفر عملية البحث عن جذر سليم، وصيغة صرفية يجوز انطباقها عليه.

تستمر بعد ذلك عمليات الفحص، وتحديد مكونات الفعل، حتى يعرض الحاسوب نتيجة التحليل الصرفي، والتي تسفر عن ثلاثة احتمالات سليمة، الاحتمال الأول: يفترض أن الفعل في الزمن الماضي، مسند إلى المفرد الغائب (هو) أو مسند إلى ضمير الغائبات (هُنَّ) ومسبوق بحرف العطف (الواو). أما الاحتمال الثاني، فيفترض أنَّ الفعل للأمر، مسند إلى المفرد المخاطب (أنتَ) أو المفردة المخاطبة (أنتِ) أو الجمع المذكر المخاطب (أنتم) في حالة التأكيد بالنون الثقيلة، والاحتمال الثالث، يفترض أن الفعل (حُزْنَ) للأمر مسند إلى الجمع المؤنث (أنتنَّ) أو أن الفعل في الزمن الماضي، مسند إلى الجمع المؤنث (هُنَّ).

١٠. نظرة تاريخية للتحليل الصرفي الآلي للغة العربية

حظيت اللغة الإنجليزية بنصيب الأسد في ميدان معالجة اللغات الطبيعية، أمّا بالنسبة للغة العربية فالأبحاث في هذا المجال انطلقت بصفة فردية في بداية السبعينيات في معامل بلدان أجنبية. فاعتمدت المحللات الصرفية العربية في بداياتها على الخبرة الفنية الحاسوبية، مع إغفال تام للخبرة اللسانية اللغوية، وربما اعتمد الحاسوبيون على بعض اللغويين التقليديين غير القادرين على فهم حاجات الحاسوب، حيث منيتْ تلك المحللات ذات الطابع الفني بفشل ذريع.

قام بعد ذلك مجموعة من الحاسوبيين واللغويين ببناء محللات صرفية آلية، تعتمد على قواعد صرفية تقوم باستخلاص عناصر بنية الكلمة، ويتميز هذا النوع من المحللات بمتانة أساسه اللغوي. فطوّر الدكتور «نبيل علي»، وأخصائية اللسانيات الحاسوبية «أمل الشامي» في عام ١٩٨٥م أول محلل صرفي آلي متعدد الأطوار[١]، وهو محلل صرفي قادر على التعامل مع أطوار التشكيل المختلفة للكلمة العربية، حيث يتكون المحلل من العناصر التالية:

١- المعالج الصرف نحوي:

يقوم هذا المعالج بدور المفكك، ليفصل جذع الكلمة عمّا يتصل به من السوابق واللواحق، كذلك يقوم برد التغيّرات الصوتية التي حدثت على عناصر بنية الكلمة الصرفية إلى أصلها؛ فعند تحليل كلمة «تمكّنّا» يدرك المعالج وجود التضعيف على النون كأحد الحالات الممكنة لتشكيل الكلمة، فيقوم بتحليلها إلى «تمكَّن + نا».

٢- المعالج الاشتقاقي:

يستخلص المعالج الاشتقاقي في طور التحليل الجذر والصيغة الصرفية من الجذع[٢] الذي فكّكه المعالج الصرف نحوي، ويتم ذلك من خلال مقارنة سلسلة حروف الجذع مع قائمة قوالب الصيغ الصرفية دون علامات تشكيلها، وبعد استخلاص الجذر يتم مقارنته بمعجم الجذور الممكنة في اللغة العربية، فإذا فشل المعالج في الوصول إلى جذر مقبول يبدأ في افتراض وجود حالة أو أكثر من الإعلال أو الإبدال حتى يصل إلى جذر مقبول.

١- يُنظر: نبيل علي، اللغة العربية والحاسوب (دراسة بحثية)، تعريب، د.ط، ١٩٨٨م، ص٣٠٨.

٢- الجذع هو الكلمة التي يمكن أن تدخلها الزائدة الصرفية، وتُشكِّل مدخلاً معجمياً في العادة.

٣- المعالج الإعرابي:

يقدِّم المعالج الاشتقاقي خرجه إلى المعالج الإعرابي، حيث يقوم بتمييز الحالة الإعرابية بناء على الوسم الإعرابي الذي تتضمنه الكلمة.

٤- معالج التشكيل:

يقوم معالج التشكيل بتحديد عناصر التشكيل الغائبة أو الناقصة، ويقوم بالرجوع إلى المعجم ليتأكد من صحة انطباق الصيغة الصرفية على الجذر رهن المعالجة.

ويتميّز المحلل الصرفي الآلي متعدد الأطوار بقدرته على فضِّ اللبس الناجم عن غياب التشكيل، أو نقصانه، وذلك باستنباطه لجميع الاحتمالات الممكنة لتحليل الكلمة، كما يتميز بقدرته على التعامل مع الأنماط اللغوية، وعناصر التشكيل، وتداخل النحو والصرف، مع اهتمامه بما يلحق الكلمة من تغيّرات صوتية ممثلة في عمليات الإعلال، والإبدال، والإدغام.

بعد ذلك طوَّر كلّ من (تيم باك والتر، وكين بيسلي) من عام ١٩٨٨م إلى عام ١٩٩٠م محللاً صرفياً لكلمات اللغة العربية، يعتمد على أسلوب التحليل الصرفي ثنائي المستوى، وفي عام ١٩٩٦م طوَّر (كين بيسلي) المحلل، واستخدم تقنية جديدة للتحليل والتوليد، حيث يقبل المحلل الكلمات العربية المشكولة كليّاً أو جزئيّاً، ويقدم عدداً من الحلول الممكنة للكلمة المحللة، ويقل عدد تلك الحلول مع استخدام علامات التشكيل.

وفي عام ١٩٩٦م قامت شركة حوسبة النص العربي في عمّان ببناء محلل صرفي عربي بهدف الوصول إلى محلل قواعدي يستخلص جذر الكلمة، ويقدِّم معلومات الضمائر والحروف المضافة إليها. كما قامت الشركة الهندسية لتطوير نظم الحاسبات (Rdi) - وهي شركة مصرية - بتطوير المحلل الصرفي Arabmorph الذي يحلل الكلمات إلى جذورها وأوزانها، ويحدد توابعها، ويضم المحلل قواعد معطيات معجمية تفصيلية لكل كلمة، ويعتمد المحلل على السياق عند تحليل الكلمات التي تحتمل أكثر من معنى.

عُقد في دمشق خلال الفترة من ٢٦ إلى ٢٨ أبريل ٢٠٠٩م اجتماع خبراء المحللات الحاسوبية الصرفية للغة العربية، وذلك بدعوة من المنظمة العربية للتربية والثقافة والعلوم، وبالتعاون مع مجمع اللغة العربية بدمشق، ومدينة الملك عبد العزيز للعلوم والتقنية بالرياض، حيث شارك في الاجتماع باحثون جامعيون عرب، وأجانب من المملكة العربية السعودية، والمغرب، والجزائر، وتونس، ومصر، وسوريا، وبريطانيا،

وفرنسا، والولايات المتحدة الأمريكية. وتمحور جدول أعمال الاجتماع حول التعريف بالمحللات الصرفية الآلية للغة العربية المقدمة من قِبَل المشاركين في الاجتماع، مع عرض نتائج تطبيق معايير تقييم المحللات الصرفية على المحللات الصرفية التي قدَّمها المشاركون. وسنعرض في عجالة بعض تلك المحللات:

١- المحلل الصرفي للغة العربية لمخبر «ميراكل» صفاقس - تونس:

يقوم محلل مخبر «ميراكل» بالتعرّف على السوابق واللواحق مع اكتشاف زوائد الكلمات، واستخراج الخصائص النحوية، والصرفية الممكنة لها. وعندما حلّلنا كلمة «وزوجناكها» باستخدام المحلل حصلنا على النتيجة التالية:

الواو	حرف عطف
زوَّجنا	فعل غير ناسخ، مسند إلى ضمير متكلم جمع مذكر / مؤنث. الصيغة: الماضي، البناء: للمعلوم، الجذر: زوج، اللاحقة: ١- (ك) المخاطب مفرد مذكر / مؤنث. ٢- (ها) الغائب مفرد مؤنث

الجدول ١: تحليل كلمة «وزوجناكها»

٢- محلل صرفي موجَّه بالتطبيقات - المعهد العالي للعلوم التطبيقية والتكنولوجيا - سوريا:

يهدف المشروع إلى تطوير محلل صرفي آلي يتمتع بالمرونة الكافية ليكون مناسباً للاستخدام في جميع المجالات. وتتكون الخوارزمية المتَّبعة في نظام المحلل من المراحل التالية:

- مرحلة تحديد نوع الكلمة: حيث يختبر المحلل كون الكلمة المدخلة أداة، أو كلمة معربة، وذلك باستخدام قائمة من الأدوات والكلمات الجامدة والمعربة.
- مرحلة الكشف عن الحروف الأصلية: وقد طوَّر المعهد خوارزمية خاصة للكشف عن الأحرف الأصلية، يصل المحلل بعد هذه المرحلة على حلول لـ٦٠٪ من الكلمات.
- مرحلة عرض الاحتمالات: يعرض المحلل مجموعة من الاحتمالات والحلول، يتضمن كل حلّ الحروف الأصلية جميعها، ولا يتضمن أي حرف زائد، مع الإشارة إلى السوابق، واللواحق الصحيحة، وإلى حالات الإدغام، والإعلال، والإبدال.
- مرحلة تصحيح الحلول: حيث يصحِّح المحلل الحلول بتطبيق مجموعة من

الخطوات الاختبارية؛ للتأكد من وجود الوزن، والجذر، وتطبيق قواعد الإبدال، والإعلال. وعند تحليل كلمة «فرق» باستخدام المحلل حصلنا على النتيجة التالية: الكلمة «فرق» الوزن: فعل، الجذر: فرق، السوابق: لا يوجد، الجذع: فرق، اللواحق: لا يوجد. وقد أغفل المحلل ذكر الكثير من المعلومات الصرفية والنحوية المهمة للفعل «فرق» كنوع الفعل، وصيغته الصرفية، وحالته الإعرابية (انظر الشكل التالي)، كما أغفل المحلل ذكر العديد من الحالات القابلة للتحليل.



الشكل ١: تحليل كلمة «فرق»

٣- محلل صرفي للغة العربية باستخدام تقنيات الذكاء الاصطناعي - فاضل سكر، وسمر معطي - سوريا:

يجرِّد المحلل الصرفي الكلمة المدخلة [١] من السوابق واللواحق، ويبحث عن الوزن الصحيح، ثم الصيغة الصرفية الصحيحة، مع إيجاد العلاقة الصرف نحوية. وقد استخدم مطورو المحلل لغة visual prolog لأنها تختلف عن اللغات التقليدية بكونها الأكثر قرباً من لغة الإنسان، وتتضمن قواعد معطيات المحلل: أوزان الأسماء، والأفعال في اللغة العربية، بحيث يسمح محرك بحث المحلل بتوليد الافتراضات بناءً على الأوزان المخزنة في قواعد المعطيات.

١- دخل المحلل يكون كلمة مشكولة كليّاً أو جزئيّاً أو غير مشكولة.

٤- محلل صرفي لكلمات اللغة العربية خارج السياق وداخله - جامعة محمد الأول وجامعة قطر:

يعمل هذه المحلل معالجة على كلمات اللغة العربية بطريقتين:

الطريقة الأولى: معالجة الكلمات خارج السياق، حيث يعمد النظام إلى تفكيك الكلمات إلى لبناتها الصرفية من سوابق، وجذوع، ولواحق؛ مما يسمح بتحديد الحلول الصرفية المحتملة لها باللجوء إلى قواعد معطيات المحلل.

الطريقة الثانية: معالجة الكلمات داخل السياق؛ ويعتمد المحلل على نموذج إحصائي يسمح بتحديد الحل الأكثر رجحاناً لكل كلمة بالنظر إلى الحلول المقترحة في الكلمات السابقة من الجملة.

٥- محلل صرفي مصدري عربي للتطبيقات العامة - مدينة الملك عبد العزيز للعلوم والتقنية:

هو محلل مصدري، وليس صرفي لكلمات اللغة العربية، يهدف إلى الوصول إلى مصدر الكلمة وليس جذرها، ويحدد لواصقها، ويعتمد المحلل التشابه في الشكل الخارجي للمفردات المولدة على الميزان الصرفي نفسه، وباللواصق نفسها. يتميز هذا المحلل بالسرعة الكبيرة، ولا يحتاج لجداول كثيرة، وهو جيد لبناء التطبيقات الحاسوبية العامة لمحركات البحث، كما يمكننا استخدامه كخطوة أولى لتفكيك المفردة العربية قبل تحليلها صرفياً.

٦- برنامج مداد للتحليل الصرفي للكلمات العربية - شركة مداد لتقنية المعلومات:

يهدف محلل مداد إلى تفكيك النصوص العربية إلى كلمات، وتحليلها لتحديد أنواع كلماتها، ثم تحديد الزوائد من سوابق ولواحق، وحروف مزيدة، بهدف الوصول إلى الجذر، مع عرض التشكيلات الممكنة للكلمة. وتضم قواعد معطيات المحلل الجذور، والكلمات الشاذة، أمّا باقي الكلمات التي تأتي حسب القواعد فيستطيع المحلل التعرف عليها دون الرجوع إلى قواعد المعطيات.

١١. طرق التحليل الصرفي الآلي[١]

هناك العديد من الدراسات التي تناولت التحليل الصرفي، وقد اتَّبعت هذه الدراسات طرقاً مختلفة لمعالجة الكلمات صرفياً، حيث يمكننا إيجازها فيما يلي:

الطريقة الأولى: طريقة قوائم الكلمات المخزنة، وتعتمد هذه الطريقة على تخزين كلمات اللغة العربية جميعها في قوائم مع مكوناتها الصرفية على شكل جداول كبيرة في قواعد معطيات ضخمة، تضم الانزياحات الصرفية بأشكالها المختلفة، ويحلِّل البرنامج الكلمة المدخلة عن طريق البحث عنها في هذه الجداول، ومن ثمَّ معرفة جذرها ببساطة، ويمكن تطبيق هذه الطريقة على نصوص معينة، مثل القرآن الكريم، أو مجموعة كتب محددة. ويعيب هذه الطريقة ضخامة حجم المواد اللغوية التي نقوم بإدخالها إلى الحاسوب، وتضييقها للتحليل اللغوي، باعتمادها مواد المعجم مرجعاً وحيداً للتحليل.

الطريقة الثانية: الطريقة اللغوية، ويكون ذلك عن طريق توصيف قواعد اللغة العربية الصرفية، وتحويلها إلى خوارزميات حاسوبية، فيحاكي المحلل عمل اللغوي عند تصريف الكلمات، وتحليلها. وتعالج هذه الطريقة عيوب الطريقة الأولى.

الطريقة الثالثة: الطريقة الرياضية، وذلك عن طريق تحليل الكلمات بشكل آلي بطريقة التجربة والخطأ والتصحيح، فالكلمة هنا مجموعة من الحروف يأخذ المحلل ثلاثة أحرف منها، ويقارنها بقائمة الجذور المخزنة في قواعد المعطيات، فإذا لم يجد المحلل الجذر في القائمة، أخذ ثلاثة أحرف أخرى، ويستمر إلى أن يجد الجذر الأقرب إلى الصواب.

الطريقة الرابعة: طريقة الأوزان، وذلك بتوليد مجموعة من القواعد النصية الآلية عن طريق المقارنة بين قائمة كبيرة من الكلمات مع ما يقابلها من مصادرها، وتُستخدم هذه القواعد لمعرفة مصدر الكلمة. فإذا انطبقت على الكلمة أكثر من قاعدة واحدة، رجَّحَ المحلل أكثر القواعد تكراراً.

١- يُنظر: عبدالله بن عبدالرحمن الزامل، العلاقة الصرفية بين الجذور والأوزان، (الأوراق البحثية للندوة الدولية الأولى عن الحاسب واللغة العربية) الرياض- السعودية، مدينة الملك عبد العزيز للعلوم والتقنية، ٢٠٠٧م، ص٢٩٩-٢٣٠.

١٢. ضوابط ومحددات للمساعدة في بناء المحللات الصرفية [١]

يعتمد النظام الصرفي والنحوي للغة العربية على مجموعة من الضوابط، والمحددات التي تُشكّل بناء كلمات اللغة العربية وجملها، وتتوزع هذه الضوابط والمحددات بين ضوابط شكليَّة، وأخرى دلالية، وهي في مجموعها محددات يمكن للعقل البشري فهمها، واستنباطها بخلاف الحاسوب الذي لا يُدرك إلا الشكلي منها، وهذه الضوابط والمحددات يمكننا ترجمتها إلى مسائل منطقية، يسهل برمجتها حاسوبياً.

إنَّ توصيف قواعد تحليل كلمات اللغة العربية مع رصد ضوابطها، يقصد إلى تمكين الحاسوب من كشف صيغ المشتقات في النصوص المشكولة، وغير المشكولة. وتنقسم ضوابط الكلمات إلى نوعين: صرفية وهي العلامات أو الخصائص التي تميِّز الكلمة من حيث بنيتها الصرفية، وما يمكن أن تقبله من تغيّرات، وما يمكن أن يطرأ عليها من أحوال. و»ضوابط نحوية تتميز بها الكلمة من خلال وجودها في التركيب، ومن خلال ارتباطها مع غيرها من الأبنية بعلاقات تحددها طبيعة التركيب نفسه، فهي ضوابط لا يتحقق وجودها إلا في التركيب«[٢].

هناك محددات، وضوابط عامة خاصة بالأسماء دون الأفعال والحروف، مثل أل التعريف، وحروف الجر، والإضافة، والتاء المربوطة، والتنوين، واتصال (ون) بجمع المذكر السالم و(ات) بجمع المؤنث السالم ...إلخ. بينما هناك محددات تميز المشتقات عن الأسماء؛ فلكل مشتق صيغة خاصة به دون غيره من المشتقات؛ فصيغة (فعّال) –مثلاً– تختص بالمبالغة، غير أن بعض صيغ المبالغة قد توافق بعض صيغ الصفة المشبهة؛ كصيغة (فعيل) ولا يفصل بينهما إلا المعنى السياقي.

١٣. مشكلات تواجه بناء محلل صرفي دقيق لكلمات اللغة العربية ونصوصها

إنَّ بناء محلل صرفي دقيق لكلمات اللغة العربية ونصوصها ليس بالأمر الهين، « بل يتطلب الكثير من الجهد، كما يحتاج إلى فرق بحث متخصصة ذات تصور كامل حاسوبياً ولغوياً، فمعظم النظم والبرامج المجرَّبة على اللغات الإنسانية لم تسلم حتى الآن من

١– يُنظر: عزت جهاد عزت العجوري، توصيف لغوي صرفي لشعر بدر شاكر السيّاب في ضوء اللسانيات الحاسوبية، رسالة ماجستير، الجامعة الهاشمية، ٢٠٠٩ م، ص ٨٧–٩٠.

٢– لطيفة النجار، دور البنية الصرفية في وصف الظاهرة النحوية وتقعيدها، دار البشير، عمّان، ط١، ١٩٩٤م، ص٤٣.

الكثير من المشكلات والصعوبات سواء على المستوى المنهجي، أو الصوري للغة»[1].
ويمكننا تقسيم هذه المشكلات إلى:

١٣ , ١ مشكلات لغوية:

إنَّ الدراسات الصرفية القديمة غير كافية لبناء محلل صرفي حاسوبي للغة العربية، فمع ما أحرزه القدماء من تقدم في دراسة صرف اللغة العربية، ومحاولاتهم الجادة في ضبط نظامها الصرفي والصوتي، «لكنها بحد ذاتها غير كافية للتعامل العلمي مع اللغة، ذلك التعامل الذي يأخذ اللغة بوصفها ظاهرة»[2]. فالحاسوب لا يتعامل إلا مع خوارزميات تضبط عمليات توليد الكلمات، وتحليلها، مما يتطلب رصد دقائق بنية صرف اللغة العربية، والإحاطة الكاملة بكلماتها. ونوجز فيما يلي أبرز تلك المشكلات اللغوية:

١- غياب التشكيل، والذي يمثل -بلا منازع- أكبر عقبة تواجه تحليل الكلمات العربية حاسوبياً؛ فنتيجة لغياب التشكيل يمكن لعدة صيغ صرفية مختلفة أن تستخدم هيكلاً واحداً للحروف، وعلى الحاسوب أن يُخَمِّنَ الصيغة الصرفية المقصودة، مثل: «كتب» يمكن أن تكون: «كَتَبَ، كُتِبَ، كُتُبٌ، كَتَّبَ». فيأخذ في الاعتبار كل حالات اللبس[3] الممكنة، وبالتالي على القائمين على بناء تطبيقات التحليل الصرفي الآلي تغطية جميع حالات اللبس التي يسببها غياب التشكيل من خلال بناء مجموعة من الخوارزميات[4] لتغطية جميع الحالات الممكنة للكلمة.

٢- الرسم الإملائي: تختلف طُرق كتابة بعض الكلمات المعرَّبة في اللغة العربية، مما يؤثر على دقة عمل المحلل الصرفي الآلي، وذلك مثل: (مسؤول - مسئول، كمبيوتر - كومبيوتر، أوروبا - أوربا)

١- عزت جهاد عزت العجوري، توصيف لغوي صرفي لشعر بدر شاكر السياب في ضوء اللسانيات الحاسوبية، رسالة ماجستير، الجامعة الهاشمية، ٢٠٠٩ م، ص ١٨-١٩.

٢- حسام الخطيب، العربية في عصر المعلوماتية - تحديات عاصفة ومواجهات متواضعة، مجلة التعريب، المركز العربي للتعريب والترجمة والنشر، العدد الثاني، ١٩٩٨م، ص٧٧.

٣- اللبس نوعان: لبس حقيقي، يكون فيه للكلمات التشكيل نفسه كما في كلمة "كَمال" فهي تحتمل: "كمال = اسم علم" و "كَمال = كـ + مال". ولبس غير حقيقي، يكون ذلك عند غياب التشكيل، كما في "كتب" حيث تحتمل "كَتَبَ و كُتُبٌ و كُتِبَ".

٤- الخوارزميات: مجموعة قواعد وقوانين مكتوبة، تستعمل لوصف الخطوات المنطقية المتبعة لمعالجة البيانات الداخلة للحصول على المعلومات والنتائج المطلوبة. وقد سميت الخوارزميات بهذا الاسم نسبة إلى العالم العربي المسلم "أبو جعفر محمد بن موسى الخوارزمي" (٨٢٥م)، والذي اشتهر في مجال الرياضيات، وقد ألف كتابه المشهور "الجبر والمقابلة".

٣- الأسماء المترجمة: هناك اختلاف في كتابة الأسماء الأجنبية باللغة العربية، مثل: (كوفي أنان - كوفي عنان، وفرانسوا أولاند - فرانسوا أولند، ومحاضير بن محمد - مهاتير بن محمد).

٤- كلمات الوقف: وهي كلمات كثيرة الورود في النصوص، ولا تحمل معاني إذا فُصلتْ عن السياق، ولا تُكوِّنُ جملة مفيدة عند استخدامها وحدها، وهي حروف وأدوات لازمة لتركيب الكلام العربي، مثل: حروف الجر، والعطف، والاستفهام، والنفي، والتعجب، والنداء، والظروف، والضمائر ...إلخ.

٥- الأخطاء الإملائية: تكثر الأخطاء الإملائية في الكتابات المعاصرة، حيث يمكننا كتابة بعض الكلمات في صور إملائية مختلفة من بينها الصورة الصحيحة، ونوجز الأخطاء الإملائية في النقاط التالية:

- الخطأ في كتابة همزتي الوصل والقطع، والهمزة المتوسطة، واختلاف كتابة الهمزة باختلاف حالة الكلمة الإعرابية، فقد لوحظ أن الهمزة تلعب دورًا كبيرًا في التمهيد للمحلل الصرفي لتحليل الكلمة المطلوبة بسهولة، في حين يشكل عليه تحليلها بدون وجود الهمزة.

- اختلاف كتابة الياء المنقوصة، مثل «قاضي» فهي منقوطة في الكتابة الشامية، وغير منقوطة في الكتابة المصرية.

إنَّ لمشكلة تفاوت رسم الكلمات جوانب سلبية على عملية تحليل الكلمات، «وترجع ظاهرة التفاوت إلى سببين، أحدهما الأخطاء البشرية، وثانيهما اختلاف الممارسات، أو القواعد المتبعة لإملاء الكلمات المعربة، والأسماء الأجنبية -كما أشرنا إلى ذلك سابقاً- والتي يختلف رسمها بين المؤلفين العرب»[1]، مما يؤثر على دقة تحليل الكلمات.

٦- من السهل على الحاسوب أن يميِّز أبنية المثنى، ولكنَّه يعجز عن تمييز كلمات أخرى، إذا عُرضتْ عليه وكانت تنتهي بألف ونون (ان) أو ياء ونون (ين)، وهي ليست مثنى، إلا إذا كان المحلل الصرفي مُزوَّداً بقاعدة معطيات تكون دليلاً إلى معرفة الكلمة بعد تجريدها من الألف والنون أو الياء والنون، «فإنْ دلَّتْ بعد التجرّد -غالباً- على

١- علي السليمان الصوينع، استرجاع المعلومات في اللغة العربية، مطبوعات مكتبة الملك فهد الوطنية، السلسلة الثانية، الرياض، ١٩٩٤م، ص٧٥.

مفردة مفيدة كانت مثنى، وإن لم تدل كانت كلمة أخرى»[1]، فكلمة (رجلان) بعد تجريدها من الألف والنون (ان) تصبح (رجل) إذن الكلمة مثنى، وذلك بخلاف كلمة (كان) فبعد تجريدها من الألف والنون تصبح (كَ) إذن هي كلمة أخرى وليست مثنى.

٧- وجود أكثر من معنى للبنية الصرفية الواحدة، ونعني بذلك أنَّ «بنية الكلمة الواحدة تحتمل أكثر من معنى واحد، فكلمة (ظهور) تكون مصدراً للفعل (ظهر) أو جمعاً للمفرد (ظَهْر). وإزالة اللبس هنا يحتاج إلى العديد من الأدلة الإضافية التي تساعد على التمييز بين الكلمتين السابقتين»[2] كأن نضع الكلمتين في سياق لغوي. كما أنه يمكن أن يكون للصيغة الصرفية الواحدة في اللغة العربية أكثر من وظيفة نحوية؛ فصيغة «فُعول» يمكن أن تكون مصدراً، نحو: «جُلوس»، وجمع كثرة، نحو: «سُيوف».

٨- التغيّرات الصوتية، وهي تغيّرات تحدث في بنية الكلمة، وتطرأ على بعض أصوات اللغة العربية في سياقات صوتية معينة، حيث تتغيّر بعض أصول الكلمة بحذفها، أو إبدالها، أو قلبها، أو إدغامها مع صوت آخر، أو إعلالها، أو إعادتها إلى أصلها؛ فالواو -مثلاً- حُذفت في «يقف» ثم عادت للظهور في «وقف»، في حين عادت ألف «جرى» إلى أصلها في «يجري». والحذف يكون في الصوامت أيضا؛ كحذف نون المثنى وياء المخاطبة إذا وقعت نون التوكيد بعدهما، كما في «يكتبانِّ» و «تكتبِنَّ»، وكذلك حذف نوني المثنى، وجمع المذكر السالم عند الإضافة، كما في «كاتبا الصحيفة، وكاتبو الصحيفة»، وكذلك حذف نون الفعل المضارع إذا كان من الأفعال الخمسة في حالة النصب، أو الجزم، أو مع نون الوقاية.

ويكون التغيير أيضاً بإبدال حرف صحيح بحرف آخر، كإبدال تاء «اِفْتَعَلَ» طاءً إذا كانت فاؤها صاداً أو ضاداً أو طاءً أو ظاءً، نحو: «اصْطَبَرَ» وأصلها «اِصْتَبَرَ»، وكذلك إبدال تاء «اِفْتَعَلَ» دالاً، إذا كانت الفاء دالاً، أو ذالاً، أو زاياً، نحو: «اِدَّهَنَ» وأصلها «اِدْتَهَنَ»، ومن التغيّر بالإبدال أيضاً، إبدال تاء «اِفْتَعَلَ» ثاءً، إذا كانت الفاء ثاءً، نحو: «اِثَّأَرَ» وأصلها «اِثْتَأَرَ».

١- محمود مصطفى عيسى خليل، إسناد الأفعال إلى الضمائر في ضوء اللسانيات الحاسوبي- ماجستير، كلية الدراسات العليا، الأردن، ٢٠١١م، ص٤٩.

٢- نهاد الموسى، العربية نحو توصيف جديد في ضوء اللسانيات الحاسوبية، المؤسسة العربية للدراسات والنشر، بيروت، ط١، ٢٠٠١م، ص٢٠٢.

ويكون التغيّر بالإعلال بالقلب، كقلب الواو والياء ألفاً، كما في «جالَ» من «جَوَلَ»، وقلب الواو ياءً، نحو: «رضي» من «رضو»، وقلب الياء واواً، نحو: «موقِن» من «مُيْقِن». كما يكون الإعلال بالحذف، نحو: «طُفْ» وأصلها «طُوفْ»، ويكون الإعلال بالتسكين، نحو: «يَسْمُوْ» وأصلها «يَسْمُوُ». أمّا التغيّر بالإدغام، فكما في «حَدَّ» وأصلها «حَدَدَ» ثمَّ حُذفت الفتحة التي بين الحرفين الثاني والثالث، مما أدى إلى الإدغام، بسبب تجاور صوتين متشابهين. وتمثّل هذه التغيّرات تحدياً يواجه الباحثين في مجال اللسانيات الحاسوبية عند تصميم المحللات الصرفية الآليّة، نظراً لكثرة هذه التغيّرات وتنوعها. مما يؤدي إلى أعباء إضافية في ردِّ الفرع إلى الأصل عند تحليل الكلمة

٩- صعوبة تعرف المحللات الصرفية الآلية على المصدر الصناعي، وذلك نحو: اشتراكيّة، انتهازيّة، شموليّة...إلخ.

١٠- اللغة العربية ذات عمليات صرفية معقدة تعتمد على العدد (مفرد، ومثنى، وجمع) والضمائر المتصلة والمنفصلة.

١١- دمج الأدوات، والضمائر المتصلة مع الكلمات في اللغة العربية، حيث تتغير صورة الكلمة في اللغة العربية عند اتصالها بالضمير، مثل: كتبت، كتبنا، كتبوا...إلخ.

١٢- عدم توفر توصيف دقيق ومتكامل لقواعد الصرف العربي، حيث تكتفي معظم كتب الصرف بشرح عام للقواعد الصرفية، مقرونة ببعض الأمثلة عن حالات الشذوذ والاطّراد.

١٣- تمثل الكلمات المركَّبة في اللغة العربية صعوبة بالغة عند تحليلها آليّاً؛ وذلك بسبب وجود الفراغ الذي يفصل بين عناصر الكلمة المركبة، حيث تختلط تلك الكلمات مع الكلمات غير المركبة.

١٣ , ٢ مشكلات حاسوبية:

هناك هوَّة كبيرة تفصل بين اللغويين والحاسوبيين، ولعل من أبرز مسبباتها «ذلك التسارع في التطور الحاسوبي من جهة، والتباطؤ في الدراسات اللغوية من جهة أخرى، إلى جانب المرجعية الغربية لتطبيقات الحاسوب، واللسانيات الحاسوبية»[1]، يضاف إلى

١- عزت جهاد عزت العجوري، توصيف لغوي صرفي لشعر بدر شاكر السيَاب في ضوء اللسانيات الحاسوبية، رسالة ماجستير، الجامعة الهاشمية، ٢٠٠٩م، ص ٢٢.

ذلك أن معظم برامج التحليل الصرفي الآلي لكلمات اللغة العربية ونصوصها هي من تطوير الحاسوبيين، حيث انشغل الحاسوبيون بالمطّرد من قواعد الصرف العربي دون النظر إلى الظواهر الشاذة، مع الاهتمام بالجانب التوليدي للكلمات دون تحليلها.

١٤. كيفية توصف القواعد الصرفية لبناء المحلل الصرفي الآلي

الحاسوب آلة صماء، لا تملك عقلاً مدركاً، ولا يمكن أن يكون الحاسوب قادراً على تقدير الأمور إلا بمقتضى حدود البرمجة؛ فهو غير قادر على تمييز كلمة (انتقل) إذا جاءت خارج سياقها، أهي فعل أمر، أم فعل ماضٍ؟ لذا ينبغي أن يوصِّف اللغوي بمساعدة الحاسوبي قواعد اللغة للحاسوب.

«وتبدأ عملية التوصيف بإيداع الحاسوب القواعد، والأساسيات الابتدائية التي يختزنها العقل الإنساني، بهدف الوصول إلى الكفاية اللغوية، ويكون ذلك عن طريق عرض منهجي قادر على استقراء القواعد، وتفصيلها وفقاً لمستويات اللغة المتفاوتة (الصوتي والصرفي والنحوي)»[١] فعند توصيف الفعل ينبغي أن نبيِّن نوعه من حيث البناء للمعلوم أو المجهول، وعلامة بنائه، وتوصيفه من ناحية صرفيّةٍ ثلاثياً أو رباعياً، مجرداً أو مزيداً، صحيحاً أو معتلاً، مع الإشارة إلى ما أصاب الفعل من إعلال أو إبدال أو إدغام، وهكذا يتم توصيف الجانب الصوتي والصرفي والنحوي.

١٥. متطلبات بناء المحلل الصرفي الآلي[٢]

أولاً- متطلبات لغوية:

١- تحديد جذور الكلمات العربية، لمعرفة أصول الكلمات التي تتشابه فيها البنية والضبط مع اختلاف الجذر.

٢- تحديد الأعلام دون تحليلها إلى مستوى الجذر.

٣- تحديد الكلمات الثابتة[٣] التي لا تُشتق منها كلمات أخرى، وهي الكلمات التي

١- نهاد الموسى، العربية نحو توصيف جديد في ضوء اللسانيات الحاسوبية، المؤسسة العربية للدراسات والنشر، بيروت، ط١، ٢٠٠١م، ص٦١.

٢- يُنظر: عبدالعزيز بن عبدالله المهيوبي، إشكاليّات تطوير محلل صرفي حاسوبي دقيق للغة العربية (محلل الخليل نموذجا)، مجلة اللغة العربية وتعليمها للناطقين بغيرها، جامعة أفريقيا العالمية، العدد ٢١، ٢٠١٦م.

٣- وتُسّمّى كلمات التوقف أو الوقف.

تثبت كما هي دون حاجة للاشتقاق منها، مثل (هؤلاء، ذلك ....).

٤- تحديد الفروق الدقيقة بين الكلمات الملبسة.

٥- بناء قاعدة معطيات للأوزان القياسية للأسماء، والأفعال المشتقة من كل جذر.

٦ - بناء قاعدة معطيات للسوابق، واللواحق، والزوائد، التي يمكن أن تأتي في بداية كل كلمة أو نهايتها.

٧- بناء قاعدة بيانات لتخزين نتائج التحليل الصرفي للكلمات.

ثانياً- متطلبات تقنية (برمجية):

١- بناء قواعد المعطيات، وبرامج إدخال المواد اللغوية وبرامج تعديلها بعد الإدخال.

٢- بناء برنامج التحليل الصرفي الآلي باستخدام إحدى لغات البرمجة.

٣- بناء برنامج لربط الجذور بمشتقاتها المختلفة الموجودة في قواعد المعطيات.

٤- بناء برنامج للتشكيل الآلي للكلمات.

٥- بناء برنامج للتصحيح الإملائي[١].

١٦ . قصور المحللات الإنجليزية عن استيعاب خصائص اللغة العربية

يواجه مطورو المحللات الصرفية العربية صعوبات تتعلق بثرائها الصوتي، والصرفي، والمعجمي الواسع، نظراً لقلة الأبحاث الأكاديمية، والتقنية المرتبطة بها، وتناثرها، وغياب التنسيق فيما بينها، سواء من الناحية النظرية أو العملية، مع قلة الإمكانيات المتاحة. «وكان من أثر ذلك أن استعار مطورو المحللات الصرفية الآلية العربية حلولاً من النظريات الخاصة باللغة الإنجليزية، حيث لم تسهم تلك الحلول في استيعاب خصائص اللغة العربية وطاقاتها حاسوبياً؛ لأنها في شتى قواعدها أشمل، وأثرى من النموذج الإنجليزي، حيث يقف المحلل الصرفي للغة الإنجليزية -نظراً لخلو اللغة الإنجليزية من خاصية الاشتقاق- عند حدود ساق الكلمة، فإذا أردنا أن نستخدمه في نطاق اللغة العربية، فلن نجد أي ارتباط بين الجذر ومشتقاته، فإذا بحثنا عن الفعل «اعلم» باستخدام محلل صرفي صُمِّمَ للغة الإنجليزية وجدناه في حرف الألف، بينما

١- يقوم المدقق الإملائي باكتشاف الأخطاء الإملائية، واقتراح التصحيحات المناسبة البديلة لها. ويُعدُّ مدقق صخر واحداً من أوائل المدققات الإملائية التجارية العربية.

نجد «تعلم» في حرف التاء، وهذا يجرِّد اللغة العربية من خاصية استدعاء الجذر لمشتقاته الذي شُيِّدتْ على أساسها ثروتها اللفظية في المعاجم، وكتب التراث المتداولة»[١].

كما يتّسم التصريف في اللغة العربية بالاطّراد التام عدا حالات نادرة، في حين يزخر تصريف الإنجليزية بحالات شذوذ متعددة. كما أن لظاهرة الإعراب أهمية كبيرة في اللغة العربية، بخلاف اللغة الإنجليزية التي تغيب عنها هذه الظاهرة بشكل شبه تام. وتتصف اللغة العربية بإمكانية دمج الضمائر والأدوات مع كلماتها، وغيابها في كلمات اللغة الإنجليزية.

ويتضح من هذا التباين أنه لا بديل من بناء نموذج لغوي لمحلل صرفي آلي مبتكر قادر على التعامل مع طبيعة اللغة العربية، يتم فيه توصيف القواعد الصرفية والنحوية بطريقة تناسب أساليب المعالجة الآلية، دون اللجوء إلى الحلول المستوردة من اللغات الأجنبية.

١٧ . لماذا تفوقت المحللات الصرفية العالمية على العربية؟

لقد أصبحت مشاريع معالجة اللغة العربية حاسوبياً سلعة تجارية تصدَّتْ لها شركات تجارية، بسبب تأخر اللغويين والحاسوبيين العرب عن الخوض في هذا المجال، عدا بعض الأعمال الفردية التي افتقدت إلى الدعم المالي. وعلى العكس من ذلك تسعى الدول المتقدمة لدعم البحوث العلمية في مجال اللسانيات الحاسوبية، مع تقديم الدعم اللازم للقطاع الخاص، ومراكز البحوث. كما أنَّ لبعثرة جهود اللغويين والحاسوبيين العرب دور كبير في هذه الفجوة الرقمية بين النظم اللغوية العالمية، والنظم العربية، فكل باحث أو شركة عربية تعمل بمعزل عن الأخرى.

١٨ . أسس مقترحة لبناء محلل صرفي دقيق للغة العربية

سنطرح هنا مجموعة من الأسس التي تهدف إلى استغلال خصائص تصريف كلمات اللغة العربية (كاطِّراد قواعد الإعلال والإبدال والإدغام)، وتتعامل مع دخائله، وتتصدى لمشاكله، وتستغل وضوحه، وتتحاور مع أوجه قصوره. ونعيد هنا لنؤكد أن الصرف العربي يمثل مجالاً نموذجياً لتزاوج الحاسوب واللغة، ونوجز هنا أهم هذه الأسس:

١- يُنظر: سعيد أحمد بيومي، أم اللغات -دراسة في خصائص اللغة العربية، والنهوض بها-، ط١، ٢٠٠٢م، ص١٠٨-١٠٩.

١- ضرورة تعامل المحلل الصرفي الآلي لكلمات اللغة العربية ونصوصها مع «أطوار التشكيل المختلفة للنصوص العربية (تامة التشكيل، والخالية من التشكيل، والمشكولة جزئياً) لذا ينبغي أن يتوافر في المحلل الصرفي الآلي الذكاء الاصطناعي الكافي؛ لتخمين النقص في عناصر التشكيل، وتغطية جميع الاحتمالات الممكنة صرفيّاً ومعجميّاً»[١].

٢- ينبغي أن يشير المحلل الصرفي إلى التغيرات الصوتية التي حدثت في الكلمة المراد تحليلها، فعند تحليل الفعل «رَدَّ» يذكر المحلل أن أصل الكلمة هو «رَدَدَ» فحُذفت حركة عين الكلمة، وأُدغمت عينها في لامها؛ بسبب تجاور صوتين متشابهين. وعند تحليل الفعل المضارع «يَرُدُّ» يذكر المحلل الصرفي أنَّ أصله «يَرْدُدُ» فحدث إعلال بنقل حركة عين الكلمة إلى فائها الساكنة، وإدغام العين باللام بسبب تجاور متشابهين. وعند تحليل الفعل «جالَ» يذكر المحلل أنَّ أصل الفعل «جَوَلَ» فقُلبت الواو ألفاً وحُذفت حركتها. أما الفعل المضارع «يَجُوْلُ»، فأصله «يَجْوُلُ»، حدث فيه إعلال بنقل حركة عين الفعل إلى فائه.

٣- أن يفرِّق المحلل الصرفي بين الصيغة الصرفية، والميزان الصرفي.

٤- أهمية التكامل بين المحللين الآليين الصرفي والنحوي، نظراً لتداخل المستويين الصرفي والنحوي، حيث يقدِّم المحلل النحوي توقعات نحوية لنوعية الكلمات، وخصائصها حسب موقعها في الجملة، ولهذه التوقعات أهمية بالغة في تسهيل عمل المحلل الصرفي عند تحليل نصوص غير مشكولة، حيث ينحصر نطاق اللبس الصرفي في حدود الاحتمالات الصرفية المقبولة نحوياً حسب مقتضيات الجملة رهن التحليل.

٥- أن يتعامل المحلل الصرفي مع جذور اللغة العربية جميعها (الثلاثية والرباعية والخماسية).

٦- فصل قواعد المعطيات[٢] المرافقة للمحلل، والقواعد الصرفية عن برنامج التحليل؛ ليكون تعديل القواعد، وتحديث قواعد المعطيات أيسر وأسهل، حيث عانت

١- نبيل علي، اللغة العربية والحاسوب (دراسة بحثية)، تعريب، د.ط، ١٩٨٨م ، ص ٢٩٩.

٢- يقصد بقاعدة المعطيات (البيانات) مجموعة من الملفات ذات الصلة ببعضها، ففي قاعدة معطيات صرفية -على سبيل المثال- يمكن أن تكون هناك عدة ملفات مترابطة مع بعضها. مثل: ملف الجذور، وملف الأفعال الثلاثية المجردة والمزيدة، وملف الأفعال الرباعية المجردة والمزيدة، والأفعال الملحقة بالرباعي...إلخ. ولعل من أهم خصائص قواعد المعطيات: الشمول، والوضوح والدقة، وقابلية التوسع والتعديل.

المحاولات الأولى لبناء محلل صرفي آلي للغة العربية من الخلط بين الجانب اللغوي، والجانب البرمجي.

٧- ضرورة تعريض المحلل الصرفي لتجارب مختلفة للتأكد من سلامة بناء خوارزميات التحليل، وكفاية قواعد المعطيات.

٨- ضرورة توفر عنصر الكفاءة والسرعة في المحلل الآلي.

٩- الالتزام بما خلص إليه البحث الصرفي الحديث من حيث اعتبار الكلمة (كتب، استخرج) - دون غيرها، هي أساس تصريف الأفعال، وجعل الجذر (ك ت ب) أساساً لعملية الاشتقاق، واستخدام الأساليب المنهجية الحديثة في صياغة القواعد الصرفية وتبويبها[١].

وأخيراً فإنَّه لبناء محلل صرفي دقيق لكلمات اللغة العربية ينبغي تطوير محلل صرفي آلي يفسر جميع مكونات كلمات اللغة انطلاقاً من المورفيمات التي تتكون منها الكلمة، مع الأخذ في الاعتبار كل الظواهر الصوتية التي تظهر عند كتابة الكلمة.

١٩ . منتهى غايتنا عند بناء محلل صرفي حاسوبي

إنَّ منتهى غايتنا عند بناء محلل صرفي حاسوبي لتحليل كلمات اللغة العربية ونصوصها هو أنْ نهيئ للحاسوب كفاية لغوية تشبه ما يكون للإنسان حين يستقبل كلمات اللغة، ويدركها، ويفهمها، ثمَّ يحللها، ويعيد توليدها. والكفاية اللغوية الحاسوبية مرتهنة بثلاثة ضوابط، هي: ١- الضابط الإملائي. ٢- الضابط الصرفي. ٣- الضابط النحوي.

١- يُنظر: علي، نبيل، اللغة العربية والحاسوب، تعريب، ١٩٨٨ م، ص٢٩٩-٢٣١.

٢٠. خاتمة

وفي ختام هذا المبحث يمكننا أن تستنتج أنَّ تطوير محللات صرفية آلية للغة العربية قد أحرز تقدماً كبيراً في السنوات الأخيرة، خاصة بعد توافر مجموعة من الأدوات التي مكَّنت اللغويين والحاسوبيين من اختبار مدى كفاية المحللات الصرفية الآلية التي طوَّرتها مراكز الأبحاث والشركات.

فكان هناك عدة محاولات لتطوير نظم آلية لتحليل كلمات اللغة العربية، بعضها يفترض وجود التشكيل الكامل، والبعض الآخر يفترض غيابه بالكامل، ومعظمها يتعامل مع الميزان الصرفي، أي الشكل النهائي للكلمة، لا مع الصيغة الصرفية؛ وذلك هروباً من مشكلات الإعلال والإبدال. غير أن الهدف بناء محلل صرفي آلي متعدد الأطوار، قادر على تحليل الكلمات المشكولة كليّاً أو جزئياً، أو غير المشكولة عن طريق تطبيق أساليب الذكاء الاصطناعي، ويعتمد على الصيغة الصرفية، ويكشف عن الانحرافات الصوتية التي حدثت في الكلمة؛ كحالات الإعلال والإبدال والتضعيف.

ويُعدُّ محلل الخليل الصرفي -الذي طوَّره مخبر البحث في الإعلاميات بجامعة محمد الأول بوجده في المملكة المغربية بالتعاون مع المنظمة العربية للتربية والثقافة والعلوم (ألكسو)، ومدينة الملك عبد العزيز للعلوم والتقنية بالمملكة العربية- بحق نقلة نوعية كبيرة جداً في مجال معالجة كلمات اللغة العربية حاسوبياً؛ حيث تمكن القائمون على هذا المحلل من جعل الحاسوب يتعامل مع اللغة بشكل يحاكي الطريقة التي يستخدمها الإنسان عند تحليل كلمات اللغة العربية.

وهنا لابدَّ من التنبيه إلى أنَّ نجاح برامج معالجة اللغات الطبيعية صرفياً يتوقف على مدى وضوح النظرية اللغوية التي يتبناها مطورو المحللات الآلية، ومدى قدرتها على تحليل الكلمات والنصوص تحليلاً صحيحاً.

المراجع

◇ **تمام حسّان**، اللغة العربية معناها ومبناها، دار الثقافة، المغرب، ١٩٩٤م.

◇ **حسام الخطيب**، العربية في عصر المعلوماتية - تحديات عاصفة ومواجهات متواضعة، مجلة التعريب، المركز العربي للتعريب والترجمة والنشر، العدد الثاني، ١٩٩٨م.

◇ **سعيد أحمد بيومي**، أم اللغات -دراسة في خصائص اللغة العربية، والنهوض بها-، ط١، ٢٠٠٢م.

◇ **عبدالعزيز بن عبدالله المهيوبي**، إشكاليّات تطوير محلل صرفي حاسوبي دقيق للغة العربية (محلل الخليل نموذجا)، مجلة اللغة العربية وتعليمها للناطقين بغيرها، جامعة أفريقيا العالمية، العدد ٢١، ٢٠١٦م.

◇ **عبدالعزيز بن عبدالله المهيوبي**، بناء خوارزمية حاسوبية لتوليد الأفعال في اللغة العربية وتصريفها - دكتوراه٬ معهد تعليم اللغة العربية - جامعة الإمام محمد بن سعود الإسلامية، ١٤٣٦هـ.

◇ **عبدالله بن عبدالرحمن الزامل**، العلاقة الصرفية بين الجذور والأوزان، (الأوراق البحثية للندوة الدولية الأولى عن الحاسب واللغة العربية) الرياض- السعودية، مدينة الملك عبد العزيز للعلوم والتقنية، ٢٠٠٧م .

◇ **عبده ذياب العجيلي**، الحاسوب واللغة العربية، منشورات جامعة اليرموك- عمادة البحث العلمي والدراسات العليا، الأردن، د ط، ١٩٩٦م.

◇ **عزت جهاد عزت العجوري**، توصيف لغوي صرفي لشعر بدر شاكر السيّاب في ضوء اللسانيات الحاسوبية، رسالة ماجستير، الجامعة الهاشمية، ٢٠٠٩م.

◇ **علي السليمان الصوينع**، استرجاع المعلومات في اللغة العربية، مطبوعات مكتبة الملك فهد الوطنية، السلسلة الثانية، الرياض، ١٩٩٤م.

◇ **لطيفة النجار**، دور البنية الصرفية في وصف الظاهرة النحوية وتقعيدها، دار البشير، عمّان، ط١، ١٩٩٤م.

◇ **مازن الوعر**، دراسات لسانية تطبيقية، دار طلاس، دمشق، ط١، ١٩٨٩م.

◇ **محمود عوض المراكبي**، تطويع تقنية المعلومات لخدمة العلوم الشرعية، السجل العلمي لندوة استخدام اللغة العربية في تقنية المعلومات، الرياض، ١٩٩٢م.

◇ **محمود مصطفى عيسى خليل**، إسناد الأفعال إلى الضمائر في ضوء اللسانيات الحاسوبي– ماجستير، كلية الدراسات العليا، الأردن، ٢٠١١م.

◇ **مروان البواب، ومحمد الطيّان**، أسلوب معالجة اللغة العربية في المعلوماتية (الكلمة – الجملة)، استخدام اللغة العربية في المعلوماتية.

◇ **مسفر محماس الدوسري**، برمجة الاسم المنسوب بياء النسب في العربية حاسوبياً – رسالة دكتوراه، جامعة اليرموك – كلية الآداب، ٢٠١٠م.

◇ **نبيل علي**، اللغة العربية والحاسوب (دراسة بحثية)، تعريب، د.ط، ١٩٨٨م.

◇ **نهاد الموسى**، العربية نحو توصيف جديد في ضوء اللسانيات الحاسوبية، المؤسسة العربية للدراسات والنشر، بيروت، ط١، ٢٠٠١م.

◇ **هدى آل طه**، النظام الصرفي للعربية في ضوء اللسانيات الحاسوبية «مثل من جمع التكسير»، رسالة دكتوراه، الجامعة الأردنية، ٢٠٠٥م.

◇ **يحيى هلال**، التحليل الصرفي للعربية، وقائع مختارة من ندوة استخدام اللغة العربية في الحاسب الآلي في الكويت، عمّان، دار الرازي.

◇◇◇

الفصل الثالث

التحليل النحوي

أ. أحمد روبي محمد عبدالرحمن[1]

ملخص البحث

تسعى هذه الدراسة إلى تقديم رؤية شاملة عن التحليل النحوي الحاسوبي في إطار تطبيقي، تحاول من خلالها الوقوف على منطلقات التحليل النحوي (التمثيل النحوي – النظرية النحوية – المحتوى النحوي) وأدواته في صورة مبسطة، بحيث تكون مدخلا مبسطًا للقارئ العربي، يمكن من خلالها فهم الصورة العامة لإطار التحليل النحوي الحاسوبي. وسعيًا لتحقيق هذه الغاية، فقد لزم الوقوف على قوام العملية النحوية/التركيبية ودورها في بناء التطبيقات الحاسوبية المختلفة التي تناظر الأداء الإنساني؛ لذا تأتي الدراسة في خمسة محاور رئيسية: تتضمن مقدمة يعرض من خلالها تأصيل طرق التوصيف النحوي، ثم عرضًا لإرهاصات التحليل النحوي الحاسوبي، ثم أهمية

١- باحث لغوي حاسوبي في إحدى شركات البرمجيات بالقاهرة – مصر. حصل على درجة الماجستير في اللغويات الحاسوبية بتقدير ممتاز من قسم علم اللغة والدراسات السامية والشرقية بجامعة الفيوم – مصر. له عدة أبحاث حول بناء المدونات المعنونة نحويًّا، وبناء قواعد البيانات الصوتية لمعالجة الكلام آليا. أنشأ بنكًا شجريًّا للغة العربية الفصحى المعاصرة. صمم محللًا نحويًّا يستند إلى طرق الذكاء الاصطناعي في المعالجة الآلية. ساهم في العديد من المشروعات التي تعنى بحوسبة اللغة العربية، منها: بناء نظم حاسوبية لتحويل الكلام العربي المنطوق إلى مكتوب، والعكس، بناء نظام حاسوبي للتشكيل الآلي، التعرف الآلي على الكينونات الاسمية. مهتم بعلم اللغة الحاسوبي، ومعالجة اللغة العربية آليًّا، وكذلك اللغويات العصبية الحاسوبية. (ahmedaruby@gmail.com)

التحليل النحوي الحاسوبي للدراسات اللغوية بصورة عامة وللغويات الحاسوبية أو معالجة اللغة الطبيعية بصورة خاصة، ويلي ذلك الخطوات الإجرائية اللازمة لبناء أية عملية تحليل نحوي حاسوبي، والتي يمكن تلخيصها في العناصر التالية على الترتيب: (النص الخام/ المدونة اللغوية - تجزئة النصوص - العنونة بالأجزاء الكلامية - الترميز بالعلاقات التركيبية)، وأخيرا تعرض الدراسة بعض موارد التحليل النحوي المتاحة للغة العربية وكذلك تطبيقاته.

◇◇◇

الفصل الثالث: التحليل النحوي

١. المقدمة

إن تزاوج الحاسوب مع غالبيَّة العلوم الإنسانيَّة لا سيما اللسانيات - بمفهومها الأعم - قد عزز من مناهجها ووسائلها، وأسرع من حركة تطورها، فضلًا عن تعاظم دورها في بناء المجتمعات الإنسانية؛ لذا أصبحت اللسانيات الحاسوبية بمستوياتها المختلفة الصوتية والصرفية والنَّحويَّة والدِّلالية بمثابة المحرك الأساسي للعديد من الأنشطة الإنسانية باعتبارها الرهان الحقيقي لاقتصاد المعرفة في ظل الانفجار المعلوماتي من جانبٍ، والدافع الأساسي للِّحاق بركب الثورة التكنولوجية من جانب آخر (روبي، ٢٠١٦م: ب).

ولما كان "النحو هو نقطةُ الالتقاء الساخنةِ بين اللسانيات والرياضيات، واللسانيات والبرمجيات باعتباره قنطرةَ الوصلِ التي تَعْبُرُ خلالها مَساراتُ الافتراضِ المتبادلِ بين علوم اللغة وعلوم الحاسب، فضلًا عن كونه المسؤولَ عن توفيرِ المعطيات اللازمة للتحليل اللغوي الأعمقِ، ألا وهو الفهمُ الآلي Automated comprehension للنصوص اللغوية" (علي، ١٩٨٨م: ٣٣٣) - فقد حاولت الدراسة وضعَ إطارٍ عام لعملية التحليل النحوي - في ضوء معالجة اللغة الطبيعية - يكون عاملًا أساسيًّا في محاولة فهم الأسس النظرية والتطبيقية لأي بناء نحوي يتوخى الطرق المنهجية.

١ , ١ التوصيف النحوي

في مضمار السعي لحل إشكالية هندسة اللغة في إطار معرفي يماهي العقل البشري كان لزامًا على الباحثين الخوض في نقل المعرفة الذهنية إلى اتساقٍ معرفيٍّ يتفق وطبيعة منحى الذكاء الاصطناعي لمعالجة اللغات الطبيعية في ضوء المناهل المعرفية الجديدة، وذلك باستخدام أدوات التوصيف المختلفة.

وقد رسم عُلَماء العربية صورة توصيفية للبنية اللغويةِ داخل عقول أبنائِها، تنطلقُ من عرضِ معطياتِ النظامِ الكلِّي عن طريقِ وصفِ الأداءِ الكلاميِّ؛ إذ كان الوصفُ باللغةِ هو الطريقةُ المثلى -آنذاك- لاستشفافِ تجلياتِ اللغةِ في العقلِ الإنسانيِّ أي: تجريدِها في عددٍ محدودٍ من القواعدِ والقوانينِ، وقد اتخذ علماءُ النحو في صيرورةِ الوصفِ مناهجَ متباينةً، تنتحي جميعُها بِعَرْضِ تجليات النظم في التركيب الجملي لمن ينشُدون تعلمَ العربيةِ فحسب (الموسى، ٢٠٠٠م: ٦١).

ثم يتوجه الوصفُ في إطار تشكل اللسانيات بِمفهومها الأعم إلى التوصيفِ والتمثيلِ اللذين يستندان إلى المنطقِ الرياضي في توصيف العموم اللغوي بغرض بناءِ نماذجَ تحاكي اللغةَ في العقل الإنساني.

١ , ١ , ١ البنية الذهنية النحوية

أجمع باحثو اللسانياتِ العصبيةِ NeuroLinguistics – من خلال التجاربِ – أن المنطقةَ اليسرى من الفَصِّ الصُّدْغِي الأماميِّ Left anterior Temporal Lobe تطبق نوعًا ما من المعالجة النحوية الأساسية (Hale & Callaway, 2014) وهذا ما يدعم مسألةَ وجودِ نمطٍ معينٍ مُنشأٍ بالدماغ البشري، إلا أنهم اختلفوا في تفسيرِ نوع هذه المعالجة، حيث ما زال يَكتنفها الغموض (روبي، ٢٠١٦م: ١٥٨).

ومع ذلك يجتهد علماء اللغة محاولين تمثيلَ تلك المعالجةِ النحويةِ الموجودةِ في الذهن البشري من خلال تصميم البناء الهندسي للفضاء الذهني وتصورِ المعنى في الدماغ الذي ينطلق من مسلمة ذهنية مُفَادُهَا "أن المعنى في اللغة الطبيعية بنيةُ معلوماتٍ مرمَّزةٍ في الذهن البشري أو هو تمثيل ذهني، ومن ثمةَ فإن المعلوماتِ التي تحملُها اللغةُ مصوغةٌ بالطريقة التي يُنَظِّمُ بها الذهنُ التجربةَ" (غاليم، ٢٠١٦م: ١٥٨).

وانطلاقا من هذا الإطار التصويري لبنية المعنى داخلَ الذهنِ، قد افترض التوليديون والتحليليون وغيرُهم من علماء اللغة أصحابِ النظريات النحوية الحديثة نموذجًا

افتراضيًّا لشكل المعرفة النحوية في ذلك الفَصِّ الصُّدْغِي، يتمثل في أن المعرفةَ النحويةَ عبارةٌ عن وحداتٍ مترابطةٍ أو متداخلةٍ تتفرّع عن بعضها البعض (1992: 251-227 Dirven, & Langacker). وتأتّى ذلك لهم مقاربةً من شكل المشتبكات العصبية synapses في الخلايا العصبية المسؤولةِ عن نقل الإشارات الكهربائية - التي تحمل المعلوماتِ - بين تلك الخلايا (شريف، ٢٠١٣م: ٥٥-٥٦).

١ , ١ , ٢ التحليل النحوي في إطار المنطق الرياضي

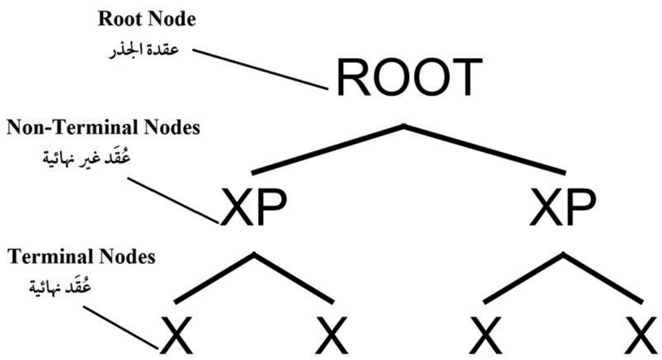
حاول العلماء تطبيق القوانينِ الرياضيةِ والمنطقيةِ لنقل أنساق هذا النموذج المعرفي الذهني إلى أنساق صورية؛ لمحاولة تَلَمُّسِ هذا العالم الخفي وإدراكِ البنية العقلية اللغوية في صورة ملموسة، فانطلقوا جميعا من مسلمة مُفَادُهَا أن النموذجَ النحويَّ الافتراضيَّ يتمثل في عدد محدود من العَلاقات والقيود، تتدرَّج فيما بينها في هيكل تنظيميٍّ أي: شجريٍّ (روبي، ٢٠١٦م: ١٥٩).

ثم أثمر تطبيق المنطق الرياضي في صوغه للنموذج النحوي الذهني تمثيلين نحويين[١]، تعددت في إطارهما العديدُ من النظرياتِ النحويةِ الحديثةِ، وهما:

- **التمثيل المكوني Constituency Representation**

"هو بنيةٌ شجريةٌ منظمةٌ، تنتظم فيها كلماتُ الجملة في شكل مركباتٍ أو مكوناتٍ نحويةٍ تبعًا لنظامٍ نحويٍّ، بحيث تظهر فيها الكلماتُ كعُقَدٍ نهائيةٍ Terminal والمركباتُ كعُقَدٍ غيرِ نهائيةٍ Non-Terminal. وهو ما يعرف بـ أشجار بنية العبارة." (روبي، ٢٠١٦م: ١٦٠). ويوضح الشكل التالي الصورة الكلية لتنظيم البنية الشجرية في إطار هذا التمثيل:

١- المقصود بالتمثيل النحوي هو تصوير بنية الجملة داخل الذهن في صورة مرئية - استنادًا إلى الأدوات الرياضية واللغوية - يمكن من خلالها تلمس مواضع الكلمات والعلاقات في أبنية الجمل.



الشكل:١ تمثيل البنية الشجرية المكونية (روبي، ٢٠١٦م: ١٦٠).

حيث X تعني الكلمات أو الوحدات، بينما X Phrase (XP) تعني المركبات أو المكونات النحوية (مركب اسمي، مركب فعلي، مركب حرفي،...).

ويتم عادة تنظيم هذه البنية الشجرية أو كتابتها عن طريق التقويس Bracketing بحيث تظهر العلاقات بينهما في صورة اعتمادية(Pustejovsky & Stubbs, 2012: 18) وقد اعتمدها عالم اللغويات الحديثة نعوم تشومسكي في تمثيله لقواعد النحو المتحرر من السياق (Jurafsky & Martin, 2007: 433) إذ هي الطريقة المثلى في تمثيل أشجار بنية العبارة حاسوبيًّا. ويمكن تمثيل الشكل السابق باستخدام هذه الطريقة في التوصيف، كما يلي:

```
(ROOT
     (XP
        (X)
        (X))
     (XP
        (X)
     (X)))
```

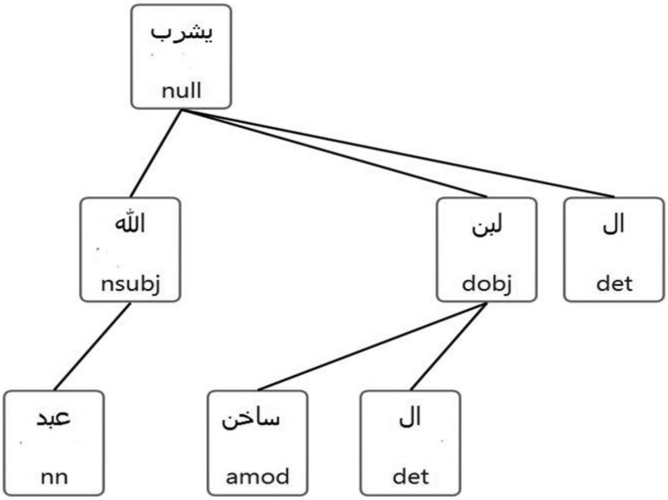
الشكل:٢ مثال لكتابة البنية الشجرية عن طريق التقويس Bracketing

والجدير بالذكر أن أصل أي تحليل نحوي يستند إلى التمثيل المكوني في التحليل، ينطلق في الأساس من التقويس، ثم يمكن تحويله – آليا – لشكل شجري – يمكن قراءته بسهولة – باستخدام العديد من الأدوات والحزم البرمجية مثل & -stanford tregex Const Tree Viewer – التي تعتمد بصورة سافرة على تنفيذ عمليات ترتيب الأقواس.

•

- **التمثيل الاعتمادي Representation Dependency**

"هو بنيةٌ شجريةٌ منظمةٌ، تنتظم فيها كلماتُ الجملة في شكل عَلاقات مكونة من مسيطرٍ وتابع أو عدة توابع، بحيث تظهر فيها الكلماتُ كعُقَدٍ نهائيةٍ Terminal فقط. وهو ما يعرف بأشجار بنية الاعتمادية"(روبي، ٢٠١٦م: ١٦٨). ويوضح الشكل التالي طريقة تنظيم الكلمات في إطار هذا التمثيل من خلال نموذج من البنيات الاعتمادية:



الشكل:٣ تمثيل البنية الشجرية الاعتمادية (روبي، أحمد، ٢٠١٦م، ص١٦٩).

يلاحظ في هذا الشكل الشجري تدرّج الوحدات من نقطة التمركز الرئيسي ألا وهي الفعل إلى نقاط فرعية أخرى تباعًا لمدى قوة الارتباط بالنقطة الأم، فنجد تدرج وحدتي «عبد الله» و «اللبن» في مرتبة مباشرة للفعل، حيث يحتلان موقعي الفاعلية والمفعولية، أما وحدة «الساخن» فهي في مرتبة مباشرة لوحدة «اللبن»، حيث إنها صفة لمفعولٍ.

ويتم تنظيم هذه البنية الشجرية أو كتابتها عن طريق تمثيلها في صورة جدولية، محددة - سالفًا - عدد الأعمدة التي تتوافق مع المعطيات اللغوية المنشودة، وذلك بغرض توحيد شكل الوثيقة من حيث المتجهات Vectors والمصفوفات Matrixes ؛ لتكون مهيئة لنظم تقنيات التعلم الآلي - فيما بعد - لبناء المحللات النحوية. ولعل أشهر طريقة تقوم على توصيف البنية الشجرية الاعتمادية وتنظيمها في صفوف وأعمدة هي هيكلة كُنل CoNLL Format التي قد تختلف من عام لآخر في تحديد عدد الأعمدة حسبما يقررها المؤتمر السنوي - لتعليم اللغة الطبيعية The Conference on Natural Language Learning (CoNLL) للآليات البرمجية - في ضوء الهدف المنشود. ويمكن تمثيل الشكل السابق باستخدام هذه الطريقة في التوصيف، كما يلي:

1	يشرب	_	_	_	_	_	0	null	_	_
2	عبد	_	_	_	_	_	1	nsubj	_	_
3	الله	_	_	_	_	_	2	nn	_	_
4	ال	_	_	_	_	_	5	det	_	_
5	لبن	_	_	_	_	_	1	dopj	_	_
6	ال	_	_	_	_	_	7	det	_	_
7	ساخن	_	_	_	_	_	5	amod	_	_

الجدول:١ تمثيل البنية الاعتمادية بهيكلة كنل CoNLL

وقد تم تمثيل هذه البنية الشجرية في صورة جدولية وفقًا للهيكلة Formatting والمعطيات اللغوية اللتين يعتمدهما محلل نوح سميث المسمى بـ (أرك)،[١] إذ يلاحظ اعتماده على الهيكلة الشائعة - في توصيف المدونات اللغوية توصيفا نحويًّا - التي تشتمل على عشرة أعمدة.

وإذا كان التمثيل المكوني ينطلق في التحليل أو الترميز من التقويس ثم بالإمكان تحويليه إلى صورة مرئية (شجرية)، فإن التمثيل الاعتمادي يستند بصورة رئيسية إلى هيكلة كنل في التحليل، مع الإمكان - أيضا - تحويلها إلى صورة شجرية باستخدام العديد من الادوات الحاسوبية مثل أداة [٢] Dependency Viewer مفتوحة المصدر للباحث الصيني جوانجشاو Guangchao Tang بجامعة نانجينغ Nanjing university الصينية(Dependency Viewer, Computer Software, 2012) .

٢. إرهاصات التحليل النحوي الحاسوبي

حظيت اللُّغة الإنجليزية دون غيرها من اللُّغات الأخرى بالسبق التُّكنولوجي، نظرًا لدروها الكبير الذي تلعبه في اقتصاد المعرفة كما أصبحت جسرًا للتَّواصُل بين فئات المجتمعات العلمية لكونها لغة الأبحاث العلمية منذ قرونٍ عديدةٍ، فلا غَرْو أن تنبت الأسس التنظيرية والتطبيقيَّة لأدوات التحليل التركيبي عبر اللُّغة الإنجليزية.

١- قامت جامعة كارنجي ميلون Carnegie Mellon برعاية هذا المشروع بقيادة الأستاذ الدكتور نوح سميث، ويمكن الاطلاع على هذا المشروع عن طريق الموقع التالي: Noah›s Ark: http://www.cs.cmu.edu/~ark

٢- يمكن تحميل هذه الأداة من خلال الموقع التالي:
http://nlp.nju.edu.cn/tanggc/tools/DependencyViewer_en.html

ومع ظهور نظرِيَّة المعلومات في بداية النِّصف الثاني من القرن العشرين على يد الأمريكيِّ كلود شانون بدأت العلاقة بين اللُّغة والإحصاء تخطو أُولَى خُطواتها نحو بناء نماذج لُغويَّة قائمة على الاحتمال الإحصائيِّ لحدُوث الظَّواهر اللُّغوِيَّة المختلفة، بقياس كَمِّيَّة المعلومات التي تتضمنها تلك المعطيات اللُّغوِيَّة. ثم كانت النَّقلة النَّوعِية على يد الرُّوسي ماركوف Markov في تأسيسه لمعالجة السلاسل الزَّمنِيَّة time series وعلاقات الارتباط بين أحداثها، والتي مكنت من تناول الظواهر اللغوية المتغيرة زمنيًّا مثل تغَيُّر الإشارة الصوتيَّة للكلام اللُّغوِيِّ، كما مكنت من إقامة نماذج إحصائِيَّة للُّغة في هيئة شبكةٍ كثيفةٍ من علاقات التَّلازُم والتَّرابُط والتَّوارُد وما شابه، وتُستخدم هذه النماذج حاليًّا في النُّظم الآلِيَّة للترجمة والفَهْرسة والتَّلخيص وفَهم النُّصوص والتحليل التركيبي (علي وحجازي، ٢٠٠٥م: ٣١٨).

ثم توالى الاهتمام من قبل الجامعات والمراكز البحثية ببناء موارد لغوية تستند إلى التوصيف النحوي؛ لاستخدامها في بناء نماذج إحصائية، يمكن من خلالها إدراك العلاقات والمتلازمات في أبنية الجمل، فكانت أول محاولة لبناء مدونة موصَّفة نحويا في عام ١٩٧٠م، حيث أعلنت جامعة لند Lund السويدية عن شروعها في بناء مدونة معنونة بالعلاقات النحوية للغة السويدية بقيادة أولف تلمان Ulf Teleman وزملائه بالجامعة نفسها؛ للوقوف على الاستخدامات النحوية لتك اللغة حينذاك (Garsid et) al., 2013: 10).

كما لم تقتصر مدونة لانكاستر-IBM للغة الإنجليزية المنطوقة على عنونة الملامح الصوتية فحسب، بل تطوّرت لتشمل العنونة النحوية، الأمر الذي يعود فيه الفضل إلى الجهود الرائدة التي قام بها اللغوي السويدي Ellegrad في عام ١٩٧٨م، وتلميذه النجيب الباحث بجامعة غوتنبرغ Göteborgs السويدية الذي عمد إلى تحليل جزء من مدونة بروان تحليلاً نحويًّا عن طريق المعالجة اليدوية الخالصة؛ إذ بدأ يتبلوّر فيها المنهج، ويتضح ملامحه، وتُرسى دعائمه (Garsid et al., 2013: 10).

وفي مطلع عام ١٩٨٠م، قرر فريق بحثي بجامعة نيجمجن Nijmegen الهولندية البدء في وضع منهجية متكاملة لعنونة المدونات اللغوية، سمّيت بـ توسكا (TOSCA Tools for Syntactic Corpus Analysis)، وذلك بهدف بناء موارد لغوية؛ للإفادة منها في الدراسات النحوية واستخدامات اللغة، وقد أفادت منها في عنونة مدونة نيجمجن (Corpus Nijmegen Kennedy, 1998: 223).

كما حاول الفريق البحثي بجامعة لانكاستر الذي أشرف على إنجاز مشروع عنْوَنة مُدوَّنة لانكاستر-أوسلو للإنجليزِيَّة البِريطانيَّة (LOB) أن يحلّل المدونة نفسها تحليلًا نحويًّا باستخدام المناهج أو الطرق الاحتمالية، إلا أن حدود إمكانات العتاد الحاسوبي Hardware (وحدة المعالجة المركزية، وحدة الذاكرة، وسائل تخزين البيانات، ملحقات الإدخال والإخراج) والإفراط في هندسة البرمجيات Software (نظم التشغيل، نظم قواعد البيانات، لغات البرمجة، نظم نقل البيانات) منعتا من إكمال هذه المهمة (Garsid et al., 2013: 10) .

وفي الفترة ما بين ١٩٨٠م إلى ١٩٩٠م، بدأت تنضج ثمار هذه المدونات الموصَّفة بوصفها موردًا لغويًّا في بناء التطبيقات الحاسوبية للغات الطبيعية، فكان من ثمارها بناء المحلّلات النحوية Syntactic Parsers التي تفيد في الترجمة الآلية وفي أنظمة السؤال والجواب Question Answering لمحركات البحث، وغيرها في متطلبات الفهم الآلي Automated comprehension للنصوص اللغوية بشكل عام.

ولما اتضح المنهج وتطوّرت أدواته، ظهرت العديد من المشروعات اللغوية الكبرى التي تتبناها المؤسسات العلمية والتجارية، وتقوم عليها فرق عمل متكاملة؛ نظرًا لضخامة العمل الذي تعجز عن تحقيقه الجهود الفردية. ففي عام ١٩٩١م، تبنَّت وحدة أبحاث حوسبة اللغة الإنجليزية (UCREL) بناء أول بنك شجري Treebank في ضوء مدوَّنة لانكاستر-IBM للغة الإنجليزية بريادة كل من روجر جارسيدي Roger Garside وجيفيري ليتش Geoffrey Leech، حيث راعت في تحليلها الاعتماد على التحليل الهيكلي Skeletal Parsing لتحديد الفئة النحوية من حيث كونها جملة كبرى sentence أو صغرى clause أوعبارة Phrase عن طريق التقويس [(١)] Bracketing؛ إلا أنها لم ترق إلى استخدام النظرية النحوية، والبنية الرياضية في تمثيلها الهيكليِّ (Gibbon et al., 2000: 33).

١- نموذج من مدونة IBM النحوية:

S[Na I_PP1A Na] [V can_MD n›t_XNOT make_VB V][N a_AT club_MM N][Tb[V pay_] [VB V] [N a_AT player_NN N][N[D so_QL much_AP D][N a_AT week_NN N]N] Tb]._. S

للاطلاع على المدونة، يمكنك زيارة الموقع التالي:

www.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/corpora/list/private/lancaster.html

وفي عام ١٩٩٥م، شرعت شبكة البيانات اللغوية بنسلفانيا Linguistic Data Consortium (LDC) في بِناء بنك شجري للغة الإنجليزِيَّة بقيادة ميتشل ماركوس Mitchell Marcus وآن تايلور Ann Taylor ، تمهيدًا لتطوير تقنيات معالجة اللغة الطبيعية، حيث وضعَ المبادئ الإرشادِيَّة لِلعَنْوَنة؛ لتكون دليلًا هاديًا للُّغوِيِّين المساعدين في تطبيقهم على نصوص المُدوَّنة، والتي كانت تتضمن تخيرهما للمحتوى النحوي وتمثيله في إطار رياضي معتمدًا على النظرية النحوية الوصفية (Descriptive theory (Abeillé, 2003: 5). وكانت هذه المرحلة أكثر نضجًا ونموًّا من المراحل أو المشروعات السابقة غير أنها تنقصها الدقة والاكتمال؛ نتيجة للغموض الحادث في التمثيل النحوي للجمل، ومع ذلك جلبت إليها الكثير من الباحثين والشركات التجارية في أنحاء العالم لاستخدام هذا المورد في أبحاثهم وتطبيقاتهم، كما مهدت الطريق أمام بناء العديد من البنوك الشجرية للعديد من اللغات الإنسانية.

ثم انطلقت العديد من المؤسسات العلمية والتجارية[١] في بناء البنوك الشجرية المختلفة للغة الإنجليزية على غرار بنك نسلفانيا الشجري، ثم توالى الشروع في مشروعات مماثلة لكثير من اللغات مثل اللغة البلغارية والصينية والتشيكية والدانماركية والألمانية والفرنسية والعربية؛ للإسهام في مجال فهم اللغة الطبيعية ومعالجتها حاسوبيًّا.

وما تزال البنوك الشجرية Treebanks أو المدوَّنات الموصَّفة نحويًّا -Parsed Tagged Corpora بعيدة عن بلوغ حد الكمال، لكنها في تطور دائم من الناحية المنهجية والنظرية في ظل ما يطرأ من نظريات لغوية حديثة ناتجة عن أثر التفاعل بين اللغة والفروع العلمية الأخرى، كما امتدت في بنائها إلى التحليل النصف آلي -Semi Automatic بعد أن كانت يدويَّة خالصة، وذلك من خلال تمريرها على المحللات النحوية الإحصائية الناتجة عن تدريبها على المدوَّنات الموصفّة نحويًّا بالفعل حسب نوع التمثيل النحوي المستخدم سواء أكان بنية العبارة أو بنية الاعتمادية.

---

١ – ومن أشهر تلك المؤسسات المعنية بالعِناية بعَنْوَنة المُدوَّنات اللُّغوِيَّة للعديد من اللُّغات: شبكة البيانات اللغوية LDC بجامعة بنسلفانيا بالولايات المُتَّحِدة الأمريكِيَّة، والمنظمة الدولية لمعاينة وثائق اللغة الإنجليزية في العصر الحديث والقرون الوسطى ICAME وجامعة نيجمجن NIJMEGEN بهولندا، وجامعة لَنْد LUND بالسُّويد، وجامعة ليدز Leeds بِبريطانْيَا.

٣. أهمية التحليل النحوي الحاسوبي

تكمن أهميَّة التحليل النحوي لِلُّغويات العامة من جانبٍ، إذ يمكن من خلالها التحقُّق من فرضيَّات النَّظرِيَّات اللُّغوِيَّة -وذلك وحده عمل تنظيري محكم يبرز طرافة المفاهيم اللغوية والنحوية- أو التَّأكُّد من فرضيَّاتٍ نحوية قائمة حول لغة معينة، فضلًا عن رسم معالم واضحةٍ لنظام الجملة في اللُّغة المَدروسة وتحديد خصائص علاقاتها التركيبية، مبيِّنةً وجوه الائتلاف والاختلاف في بنية الجملة (عاشور، ١٩٩٢م: ٨) وهذا ممَّا يسهّل النَّظر في الغايات التَّربوِيَّة في تعلُّم اللُّغة ومعرفة خصائصها، كما يستطيع المُعلّم والمُتعلّم على حدٍّ سواء -من خلال التَّنقيب أو البحث في المدوَّنات المحللة تركيبيًّا- معرفة السِّمات اللُّغوية لكلمةٍ معينةٍ وسياقها في الجملة واختلاف معانيها اعتمادًا على السِّياق والموضوع (Sharaf & Atwell, 2010: 2-3).

وللُّغوِيات الحاسوبية من جانبٍ آخرَ، فهي قوام تطبيقات معالجة اللغة الطبيعية التي تناظر الأداء الإنساني، والتي يمكن تلخيصها في عنصرين أساسيين:

- **المحللات النحوية**

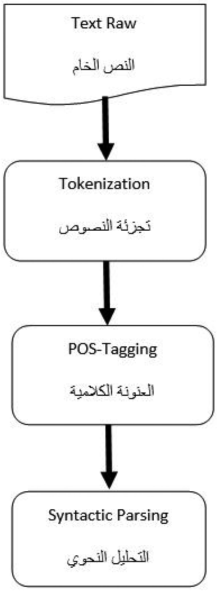
تمثل تطبيقات التحليل النحوي (المحللات النحوية) صلب العديد من تطبيقات معالجة اللغة الطبيعية المختلفة مثل: الترجمة الآلية، إجابة الأسئلة، التدقيق النحوي، البحث الدلالي، التخاطب مع قواعد البيانات باللغة الطبيعية، فهم اللغة الطبيعية، وغيرها.

- **استخلاص المعلومات**

تزايدت المعلومات بصورة متسارعة في ظل تنامي الثَّورة المعلوماتيَّة العارمة، وانتشار الشَّبكة المعلوماتيَّة التي تتعامل مع العديد من المعارف مثل الكتب والدَّوريَّات والنَّشرات والصُّحف والأبحاث العلميَّة، وغيرها من دوائر المعارف الأخرى؛ مما أدَّى إلى صعوبات جمَّة في استخلاصها وتنقيتها من التَّلوُّث المعلوماتيِّ، فكان لزامًا على المعلوماتيين مواجهة هذه التَّحديَّات ببناء نظم برمجية قادرة على استخلاص المعلومات وتجميعها، تنطلق من المعطيات اللُّغويَّة والتركيبية؛ للكشف عن محتوى النَّصِّ(روبي، ٢٠١٦م: ٨٢).

٤ . خطوات التحليل النحوي الحاسوبي

تقتضي أية عملية تحليل نحوي الاستناد إلى عدة طرق إجرائية متدرجة، تنطلق من تحديد النص الخام ثم تقسيمه إلى أجزاء على مستوى الجملة والوحدات، وصولًا إلى تعيين المعلومات الابتدائية الممثلة في التعرف على أنواع الأجزاء الكلامية التي تساعد بدورها في تعيين العلاقات التركيبية. ويوضح الشكل التالي خطوات التحليل النحوي:



الشكل: ٤ خطوات عملية التحليل النحوي.

٤ , ١ النص الخام / المدونة اللغوية Corpus

يمكن تعريف المدونة اللغوية بأنها كتلة غير منتظمة من النصوص المكتوبة أو المنطوقة، يمكن التعامل معها آليًّا والتحكم في بياناتها ومدخلاتها بالإضافة أو الحذف أو التعديل من خلال محررات النصوص (السعيد، ٢٠٠٨م: ١).

واختيار نصوص المدونة اللغوية المعنية بالتحليل يعتمد على عدة معايير يجب مراعاتها في ضوء الهدف المنشود من التحليل، منها:

- انتماء النصوص لأي مستوى (الفصيح - المعاصر - العامي).
- مصدر النصوص (الكتب - الصحف الورقية - الصحف الإلكترونية).

- طريقة اختيار النصوص (الحصر الشامل - العينات الإحصائية).
- تنوع النصوص أو التركيز على نوع محدد من النصوص.

٤ , ٢ تجزئة النصوص Tokenization

«يعتقد الباحثون في العلوم المعرفية أن إدراك المخ البشري للنص اللغوي وفهمه يتم من خلال تقسيمه إلى مكونات أو وحدات منفصلة، ثم تنظيمها بطريقة متسلسلة، بحيث تعمل الواحدة تلو الأخرى في الدماغ البشري»(روبي، ٢٠١٦م: ٤٧).

ومما يؤكد هذا الزعم أن تحليل النص لغويًّا يتطلب مرتكزًا أساسيًّا وهو تفكيك الوحدة النصية إلى مكوناتها الجزئية، بحيث تتيح لنا معرفة بنياتها الداخلية، وإن توخّي التحليل النحوي يتصل بتحليل الجملة إلى عناصرها الأولية للكشف عن علائقها وربط مكوناتها (روبي، ٢٠١٦م: ٤٧).

ويشير مفهوم تجزئة النصوص إلى تقسيم النص-آليا- إلى وحدات منفصلة من خلال جملة من المعطيات اللغوية اللازمة التي تكون دليلا مستأنسًا للحاسوب في تعيينه إلى هذه الوحدات (Attia, 2007: 65).

إذا كانت تجزئة النصوص هي العمود الفقري لتطبيقات معالجة اللغة الطبيعية، فإن دقة هذه التجزئة تنعكس على أداء التطبيقات اللغوية، كما اتضح في الشكل السابق. وتتمُّ عملية تجزئة النُّصوص -بالنسبة للُّغة العربيَّة- على ثلاثة مستويات:

٤ , ٢ , ١ التجزئة على مستوى الجملة

تتصل تجزئة الجملة بشكل عام بالتعرّف على معيار تحديد أبعاد الجملة الذي يمثله مقياس الشكل النحوي أو المعنى التام. ويتم التقسيم في تحديد أبعاد الجملة - حسب الإسناد والتركيب التام المفيد، وما بين الجمل من علاقات الربط بواسطة أدوات الاستئناف والعطف (روبي، أحمد، ٢٠١٦م، ص٤٧).

وتتخذ آلية تجزئة النصوص Tokenizer من علامات الترقيم وسيلة لتجزئة النص إلى جمل منفصلة (Habash & Faraj,125 :2009). إلا أنَّ هذه الآلية تحتاج إلى تتمة الأدلة ليكتمل فيها من المحددات التي تبلغ حد الكفاية في تجزئة النص إلى جمل؛ لما في علامات الترقيم من لبس يشوبها، فقد تعددت وظائفها في النص الكتابي بين وظيفتها الأساسية وما يتفرع منها، فعلى سبيل المثال لا الحصر النقطة التي تستخدم كمحدد

للدلالة على نهاية الجملة، تستخدم بين الاختصارات مثل أ.د، ص.د.ب، وغيرها، وكذلك الفاصلة التي تعد ملمحًا مميزًا للفصل بين الوحدات أو المكونات في الجملة، تستخدم حال الأرقام العشرية (روبي، ٢٠١٦م: ٤٩).

٤ ,٢ ,٢ التجزئة على مستوى الوحدات/ العناصر الرئيسية

العنصر اللغوي Token هو أصغر وحدة نحوية، يمكن أن تكون كلمة أو جزءًا من الكلمة، أو تعبيرًا اصطلاحيًّا، أو مركبًا أو رمزًا (Attia, 2007: 65)، ومادامت العناصر اللغوية الرئيسية هي الجزء الملموس من التحليل فيمكن أن نطلق عليها أيضا «وحدات التحليل النحوي»(شمس الدين، ١٩٩٥م: ٦٨).

والوحدة الرئيسة هي البناء اللغوي المتكامل سواء أكانت كلمة أو علامة أو رقمًا، وتعد عنصرًا أساسيا في النص اللغوي. و تشتمل التجزئة على مستوى الوحدات أو العناصر الرئيسية Main Tokens على ثلاثة مستويات:

أ- الكلمة

تعرَّف الكلمة في اصطلاح اللغويين بأنها «صيغة ذات وظيفة لغوية معينة في تركيب الجملة، تقوم بدور وحدة من وحدات المعجم، وتصلح لأن تفرد، أو تحذف، أو تحشى، أو يغير موضعها، أو يستبدل بها غيرها في السياق، وترجع مادتها غالبا إلى أصول ثلاثة»(حسان، ١٩٥٥م: ٢٦٢).

أمَّا في عرف الحاسوب، فهي حيّز من الحروف المتشابكة، أو الحروف المفردة أو العلامات، أو الرموز، يحيطه من جانبيه مساحات بيضاء White Spaces. وهذه المساحات هي المعطيات التي تفضي إلى حدود الكلمة لتجزئة الوحدات الرئيسية في النص(السعيد، ٢٠١١م: 84؛ Jurafsky, & Martin, 2007: 26).

وثمة عديد من آليات تجزئة النصوص إلى وحدات رئيسية - وغالبًا هذه الآليات يتم إدراجها في المحللات النحوية - منها أداة التجزئة العربية Arabic Tokenizer المدرجة في محلل ستانفورد التركيبي.

ب-المركب غير الكلامي

«هو انضمام كلمة إلى كلمة فأكثر، وتكون بحكم المفرد نحويا ودلاليًّا»(الدحداح، ٢٠٠٠م: ٢٩٤) مثل: عبد الله، جاب الله، أبو عيد، إسلام أون لاين، الصهيو أمريكي، الجيو إستراتيجية.

ويتم معالجة المركب غير الكلامي في النص اللغوي قبل إجراء عملية التجزئة من خلال وضع علامة الشرطة (-) بين الكلمة الأولى والكلمة الثانية؛ ليكونا في حكم الكلمة الواحدة. مثال ذلك : عبد-الله، جاد-الله.

ج- الرمز أو العلامة

يشمل جميع الرموز المستخدمة في النص العربي، مثل علامات الترقيم والأرقام، وغيرها من الرموز.

٤ ,٢ ,٣ التجزئة على مستوى الوحدات/ العناصر الفرعية

يمكن أن نعرّف العنصر اللغوي أيضا بأنه «بناء لغوي يحدده مستوى التحليل»(شمس الدين، ١٩٩٥م: ٦٩)، إذ نجد أن العنصر اللغوي الرئيسي قد يكون مكونًا من مورفيم/ عنصر فرعي واحد أو أكثر من مورفيم، فعلى سبيل المثال يمكن للكلمة المفردة (العنصر الرئيسي) أن تشمل أربع وحدات فرعية سواء أكانت سوابق أو لواحق (Attia, 2007: 65).

وتتوقف حدود عملية تجزئة العناصر الرئيسية إلى عناصر فرعية على طبيعة الغرض من البحث، أي ما العناصر الفرعية المراد تجزئتها من العناصر الرئيسية؟ ويقتضي لتحليل الجملة العربية تجزئة عناصرها الأساسية التي تكوّن العلاقات النحوية في بنية الجملة (روبي، ٢٠١٦م: ٦٠).

ولما كان الكلام سلسلة من الجزئيات المتتابعة، كان لزامًا على تلك الدراسة أن تعرض أنوع تلك الجزئيات:

ثمة أنواع من المورفيمات اللصقية Concatenative Morphemes في اللغة العربية: الجذع (Stem) واللواصق (affixes) والزوائد (Clitics) (Clitics Jurafsky & Martin, 2007: 7).

أ- الجذع Stem : هو جزء أساسي من الكلمة، يأتي مشتقًا أو جامدًا، وينتج عن اتحاد المورفيمات اللصقية للكلمة، ومن أمثلته: الجذع (كتب) الذي تكون عنه التركيب في(وسيكتبونها) والجذع (مكتب) في صيغة الجمع (المكتبات).

ب- اللواصق Affixes: هي مورفيمات تتعلق بجذع الكلمة، وهناك نوعان من اللواصق:

١) السوابق (Prefixes): والسابقة مورفيم يسبق الجذع في أوله، ومن أمثلته: نون في الفعل المضارع في "نفعل- نعمل- نشكر".

٢) اللواحق (Suffixes): واللاحقة مورفيم يلحق الجذع في آخره، ومن أمثلته: الواو والنون في جمع المذكر السالم في" المسلمون-العاملون".

ج- الزوائد Clitics: هي مورفيمات نحوية تكون مقيدة بكلمات أخرى، وتتعلق بجذع الكلمة بعد اللواصق. وهناك نوعان من الزوائد (Attia, 2007: 65):

١) الزوائد في بداية الكلمة (Proclitics): فهي تشبه اللواصق، ولكنها تختلف اختلافًا واضحا عن اللواصق التي تمثل جزءًا من الكلمة صوتيًّا وبنيويًّا، ومن أمثلتها: حروف العطف، وحروف الجر، والنداء.

٢) الزوائد في نهاية الكلمة (Enclitics): وهي التي تعقب الكلمة، مثل الضمائر المتصلة.

وهناك العديد من الأدوات الحاسوبية - مفتوحة المصدر- التي تعمل على تجزئة الوحدات الفرعية في النص أشهرها أداة MADAMIRA التي تم تطويرها من قبَل فريق معالجة اللغات الطبيعية بمركز أنظمة التعلم الحاسوبي بجامعة كولومبيا CCLS.

٤, ٣ العنونة بالأجزاء الكلامية POS Tagging

هي عمليَّة تعيين الأجزاء الكلامِيَّة وما تحمله من سمات صرف-نحوِيَّة لكلِّ كلمة منفردة بمعزلٍ من سياقها الإعرابيِّ في النَّصِّ، وذلك بإلحاق كل مفردة برمز Tag أو عدَّة رموز تعبِّر عن الجزء الكلاميِّ وما يحتويه من مورفيمات أخرى.(Van, 1999: 3; Attiya, 2004: 42) مثال ذلك: وقع VB/ الاختيار NN/ .

ويقتضي التَّوصيف أو العنونة منهجًا يستند إلى مبادئ نظرية تسوّغ التَّحليل والتَّأويل بالاعتماد على الشمول والاختصار في اختيار مجموعة من المعطيات اللغوية Tags set تمثل تلكم المعلومات المنشودة، وعدم التَّناقض في التَّحلِيل بالتعرّف على النظائر في ضوء المعطيات (المقولات) اللغوية (Kennedy, 1998: 220).

وقد تعدَّدت منهجيات الأجزاء الكلاميَّة للُّغة العربيَّة التي تُصنّف مفردات النص في ضوء وصف الواقع اللغوي، منها:

١. فئة خوجة الكلامية
٢. فئة باكولتر الكلامية

٣. فئة بييز الكلامية
٤. فئة آر دي آي الكلامية
٥. فئة القريني الكلامية
٦. فئة كاليك الكلامية
٧. الفئة الكلامية للنص القرآني
٨. فئة كاتب الكلامية

٤ , ٤ الترميز بالعلاقات التركيبية Syntactic annotation

يتوخى الترميز بالعلاقات التركيبية عدة طرق إجرائية، لا تنفك إحداهن عن الأخرى، فهي بمثابة أجزاء اللوحة التشكيلية التي لا يكتمل معناها إلا إذا اتحدت وتكاملت مع بعضها البعض (Rambow, 2010). وهذه الطرق:

٤ , ٤ , ١ التمثيل النحوي Syntactic Representation

ويمكن تعريف التمثيل النحوي بأنه النموذج الرياضي الذي يعرض بنية الجملة بشكل تصويري في إطار النظرية النحوية والمحتوى النحوي. وقد أسهم هذا التمثيل النحوي في توضيح طبيعة المعرفة وأنساقها والفهم والتأويل، وفي التقدّم التقني للحوسبة computation (الفهري، ١٩٩٠م: ١٧).

وهناك نوعان من التمثيل النحوي، تعددت في إطارهما العديد من النظريات النحوية أو الصورنة النحوية:

- **التمثيل المكوني Constituency Representation**

تعددت النظم أو النماذج الرياضية التي تصور بنية الجملة إطار مكوني، حيث قدم تشومسكي في كتابه التراكيب النحوية عام ١٩٥٧م نموذجًا رياضيًّا يسمى بالنحو المتحرر من السياق Context-Free Grammars [1] - وهو النموذج الأكثر شهرة- يصف بنية الجملة استنادًا إلى عدد من القوانين التي تعبر عن أركان الجملة المتمثلة في الفئات الرئيسية (الاسم، الفعل، الصفة،...)، والمركبات (مركب اسمي، مركب

١- وهناك العديد من الصوريات النحوية Formalisms التي انبثقت عن النحو المتحرر من السياق، منها: نحو بنية المقولات النحوية العامة (Generalized Phrase Structure Grammar (GPSG، نحو بنية المقولات النحوية المعتمد لعلى الرأس (Head driven Phrase Structure Grammar (HPSG، والنحو الوظيفي المعجمي Lexical Functional Grammar (LFG، والنحو المقولي (Categorical Grammar (CG.

فعلي،...) التي قد تكون مزيجًا من الفئات النحوية المتتالية أو تكون فئة نحوية واحدة. ويتم صياغة هذه القوانين في صورة هذه المعادلة:

$$X \rightarrow Y$$

حيث الرمز X يشير إلى العنصر المفرد single element، أما الرمز Y فيشير إلى سلسلة مكونة من عنصر أو أكثر، وتتضح الصورة بالنظر للأمثلة التالية:

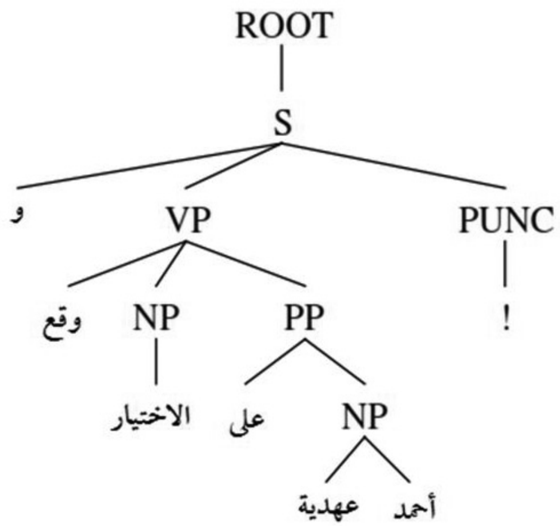
$$VP \rightarrow V+NP\ (PP)\ (Adj)$$

$$NP \rightarrow N+(NP)$$

$$PP \rightarrow P+NP$$

وهذه القوانين وحدها لايمكن أن تصف أبنية الجملة، إذ لابد من إطارٍ نظري يمكن من خلاله تنظيم أو إحكام البنية التركيبية في الجملة. كما لا يمكن أن يقدم الإطار النظري دون هذه القوانين أو التمثيل (Rambow, 2010). وسنتناول الإطار النظري (النظرية النحوية) بالتفصيل في الخطوة الثالثة من خطوات التحليل النحوي.

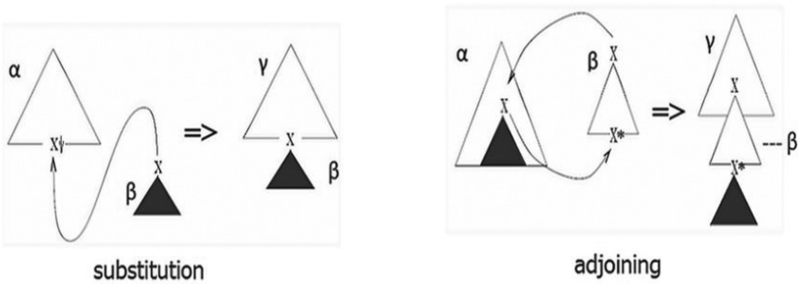
ويوضح الشكل التالي تمثيل بنية المكونية في إطار نموذج النحو المتحرر من السياق:



الشكل:٥ التمثيل المكوني في إطار نموذج النحو المتحرر من السياق. [١]

وقدم أرفيند جوشي Aravind Joshi عام ١٩٨٥م نموذجًا رياضيًّا آخر يسمى بالنحو الأقل ارتباطًا بالسياق Mildly Context-Sensitive Grammars، ينطلق من النحو المتحرر من السياق إلا أنه يستند إلى الأشجار المتجاورة – بدلًا من قوانين بنية العبارة – في صياغته (Abeillé, 2000: 19). وهناك العديد من الصوريات Formalisms التي تنبثق من هذا النموذج، أشهرها نحو الأشجارة المتجاورة Tree Adjoining (TAG) Grammars الذي يصف بنية الجملة استنادًا إلى عدد من القوانين التي تعتمد على الاستبدال أو الإحلال والتجاور في أبنية الأشجار (Schmidt, 2005: 3). ويوضح الشكل التالي الصورة العامة لقوانين الاستبدال والتجاور في أبنية الأشجار:

١- يستند هذا المثال إلى النظرية الوصفية في تنظيم أبنية العوامل، معتمدًا على المعلومات أو المحتوى النحوي العام للمكونات (NP المركب الاسمي، VP المركب الفعلي، PP المركب الحرفي).

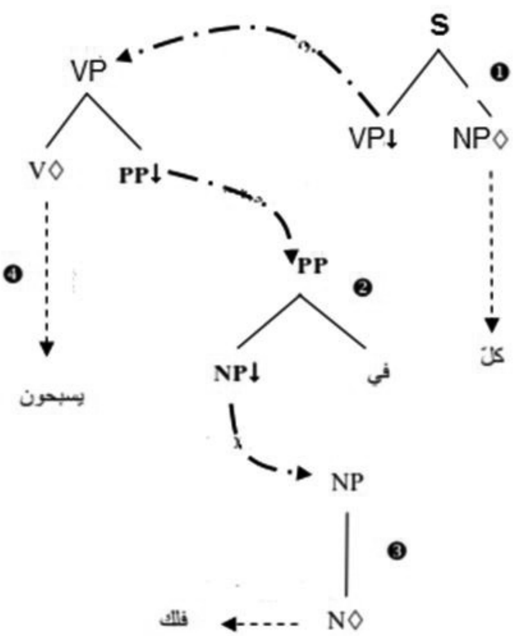


الشكل:٦ الاستبدال والتجاور في نحو أشجار التجاور (Abeillé, 2000: 19).[١]

يهدف هذا النموذج الرياضي إلى تحليل أو تكوين الشجرة التحليلية للجملة في صورة تسلسلية من الشجيرات استنادًا لعمليتي الاستبدال Substitution والتجاور Adjoining اللتين تحكمهما العلاقات - التي تبرزها النظرية النحوية - بين أجزاء الشجيرات، إذ يتم استبدال الشجيرة الصغيرة بشجيرة أكبر في الجملة وهكذا الحال إذا كانت الشجيرة الصغيرة جزءا من شجيرة أكبر، أما إذا كانت الشجيرة هي بنية وصفية لشجرة أخرى، فيتم وضعها بالتجاور إلى أن يتم تكوين الشجرة النهائية للجملة.

ويعرض الشكل التالي تمثيل البنية المكونية في إطار نموذج النحو الأقل ارتباطًا بالسياق (نحو الأشجار المتجاورة):

١- هذا الرمز(↓) يدل على الإحلال أو التبادل substitution أما الرمز الآخر (*) يُرمز لعقدة القدم في حالة التجاور adjoining.

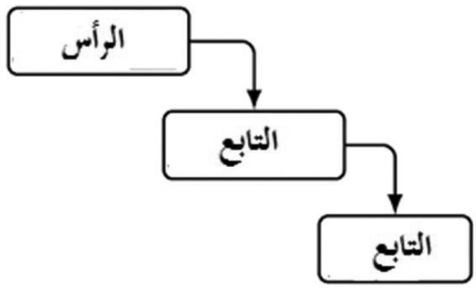


الشكل:٧ التمثيل المكوني في إطار نموذج النحو الأقل ارتباطًا بالسياق(نحو الأشجار المتجاورة) (Fraj, et al., 2008).[1]

- **التمثيل الاعتمادي Representation Dependency**

قدم اللغوي الفرنسي لوسيان تنير(L.Tesnière) في كتابه عناصر النحو التركيبي عام ١٩٥٩م نموذجًا رياضيًّا ينطلق في تصوير بنية الجملة – في إطار اعتمادي – من نقطة التمركز الرأس ثم التابع ثم ما يتبع التابع وهكذا (البحيري، ١٩٨٨م: ١٢). ويوضح الشكل التالي صورة هيكلة تمثيل البنية الاعتمادية:

١– يمكن تحميل هذه الأداة من خلال الموقع التالي:
http://nlp.nju.edu.cn/tanggc/tools/DependencyViewer_en.html

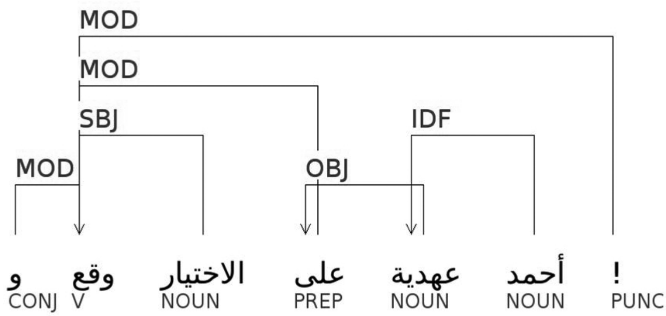


الشكل:٨ هيكلة تمثيل البنية الاعتمادية.

ويقتضي هذا التتابع أو التسلسل المتدرج في صورة اعتمادية إطارًا نظريًا، أي نظرية نحوية سواء أكانت وصفية أو تفسيرية، يمكن من خلالها إحكام البنية التركيبية في الجملة عن طريق العلاقات بين الكلمات.

وهناك العديد من الصوريات التي تنطلق من هذا النموذج الرياضي منها: نحو الكلمة Word Grammar لهدسون عام ١٩٨٤م، ونص المعنى Mining Text لملكوك عام ١٩٨٩م، والنحو الارتباطي Link Grammar لسليتور عام ١٩٩٣م.

يعرض الشكل التالي نموذجًا للتمثيل الاعتمادي في إطار نحو تنيير، وهو ما يقارب فكرة المسند والمسند إليه في النحو العربي:



الشكل:٩ التمثيل الاعتمادي في إطار نموذج تنيير الرياضي. [١]

١ – يستند هذا التحليل إلى معطيات النظرية الوصفية، ممثلا بالمعلومات التي أنتهجها بنك كولومبيا الشجري.

٤ , ٤ , ٢ المحتوى النحوي Syntactic Content

يشير مصطلح المحتوى النحوي إلى المعطيات النحوية التي أنتجتها لغة الوصف Meta Language من خلال التجريد والتوصيف للظاهرة اللغوية؛ لتشمل المقولات والمكونات والرموز التي تعبّر عن القواعد والعلاقات النحوية، والتي يتعذر بدونها تمثيل النظرية النحوية، كما لا يمكن – أيضًا – أن تمثَّل وحدها دون نظرية نحوية، لكن يمكن أن نشير إليها فحسب أثناء العملية التعليمية (Rambow, 2010).

وتتقيد المدونة اللغوية المعنية بالتحليل أو بالوصف للعلاقات التركيبية بعدد محدد من المعطيات النحوية Tag set التي تخضع للهدف المنشود من البناء، كما تتقيد المعطيات بنوع التمثيل المتوخى في التحليل، إذ يقتضي التمثيل المكوني معطياتٍ حول الوحدات المكونة للجملة التي لا يتطلبها التمثيل الاعتمادي مثل (نوع المركب، نوع الجمل، نوع الوحدات المكونة، الإشارات المشتركة Co-indexing، الفصائل المحذوفة) ويقتضي التمثيل الاعتمادي بيان العلاقات النحوية الممثلة (الفاعلية، المفعولية،...) التي قد لا يقتضيها التمثيل المكوني.

٤ , ٤ , ٣ النظرية النحوية The syntactic theory

النظرية هي "مجموعة متجانسة من المبادئ والأسس والقواعد، التي ينتظمها مفهوم مستوعب للكلام وأحكامه، قادر على إيصال الدلالات الصحيحة للمخاطب"(عفيفي، ٢٠٠٣م: ١٩٢) وتنطلق هذه المبادئ من تمثيل المعنى في إطار المعطيات النحوية التي تعدّ خصائص عامة في كثير من اللغات أو في اللغة المعينة.

وتجدر الإشارة إلى أن اختيار نوع التمثيل النحوي يحدد اتجاهات النظرية المتوخاة ما بين النظام التحليلي أو التوليدي، حيث يتلاءم النظام التحليلي مع التمثيل الاعتمادي، بينما يتلاءم النظام التوليدي مع التمثيل المكوني (روبي، ٢٠١٦م: ١٧٩).

وتدور النظريات النحوية في فلكين أساسين، فَلَك يعنى بمبادئ الوصف النحوي للغةٍ مخصوصةٍ، وفلك آخر يعنى بالمبادئ التفسيرية للظاهرة النحوية في جميع اللغات الإنسانية. أو بقول آخر، تدور النظريات النحوية فيما أرساه دي سوسير بجعل اللغة ظاهرة زمكانية يمكن وصفها وصفًا آنيًّا، وفيما افترضه تشومسكي حول النحو الكلي UG بأنه موجود في أذهان الأطفال منذ الولادة، ويتمثل في مجموعة من المبادئ والمحددات (روبي، ٢٠١٦م: ١٧٩).

ومن ثم تبلورت النظريات النحوية – في إطار التمثيل النحوي – في نوعين:

١- **النظرية الوصفية:** هي ضرب من بيان مبادئ طرق انعقاد الربط والارتباط بصورة محكمة في الجملة، وكذلك بيان موقع المكونات أو الوحدات في الجملة (روبي، ٢٠١٦م: ١٩٢).

وقد تم تطبيقها في إطار التمثيل المكوني في البنك الشجري العربي Arabic Treebank وفي إطار التمثيل الاعتمادي في بنك كولومبيا الشجري CATIB.

٢-**النظرية التفسيرية:** «هي مجموعة المبادئ المنظمة التي ينبغي أن يلحظها البحث اللساني من حيث هي مشتركة بين اللغات وتلتزم بها اللغات»(زكريا، ١٩٨٦م: ٧٧). وهي ما تسمى بالقواعد الكلِّيَّة أو النحو الكلي الذي يقوم على المبادئ العامة General Principles المشتركة بين أنحاء اللغات.

وقد تم تطبيقها في إطار التمثيل المكوني في البنك الشجري للغة الصينية CTB، حيث اعتمد على نظرية السين البارية X-bar كأساس نظري، يمكن من خلاله تنظيم القوالب في ضوء الأحكام التركيبية. أما تطبيقها في إطار التمثيل الاعتمادي فقد تم تطبيقها من قبل جوكايم نيفر، وريان ماكدونالد وغيرهم في بناء مشروع الاعتماديات العالمية Universal Dependencies (UD) لتوصيف اللغات الإنسانية بمنهجية اعتمادية واحدة، وذلك استنادًا للنظرية الاعتمادية العامة التي وضعتها ماري مرنف أستاذة اللغويات الحاسوبية بجامعة ولاية أهاويو.

٥. موارد التحليل التركيبي للغة العربية وتطبيقاته

على الرغم من أن هناك فقرًا شديدًا في توفّر الموارد اللغوية الموصفة للغة العربية، فإن هناك تقدمًا ملحوظًا في توصيف المدونات العربية توصيفًا تركيبيًّا، ومن ثم بناء النماذج الإحصائية للغة لإنتاج العديد من تطبيقات التحليل النحوي، والترجمة الآلية، وإجابة الأسئلة.

ولعل أشهر مدونة نحوية للغة العربية هي تلك التي أنتجتها مؤسسة شبكة البيانات اللغوية LDC ببنسلفانيا وهي مدونة بنك بنسلفانيا الشجري (PATB) ، تليها مدونة بنك براغ الاعتمادي (PADB)، إذ تتشابهان نسبيًّا في مدى ثراء المعلومات اللغوية المقدمة، مع الاختلاف الواضح فيما بينهما في تمثيل تلكم المعلومات، فضلًا عن الالتقاء في الغرض المنشود من البناء ألا وهو بناء محلل نحوي.

ثم قدم مؤخرًا مركز أنظمة التعلم الحاسوبي (Center For Computational Learning Systems) بجامعة كولومبيا بنكًا شجريًّا نحويًّا ثالثًا (CATiB)، ينطلق من تقليص حجم المعلومات اللغويات المقدمة في المدونتين السابقتين، ومحاولة تجنب المعلومات التي لا فائدة منها بغية تسريع عملية الترميز Annotation.

ثم تعددت المحاولات الفردية للباحثين في عنونة بعض المدونات اللغوية صغيرة الحجم، بغرض تقديم الأطروحات العلمية والدراسات البحثية، منها: محاولة الباحث في أطروحته التي تقدم بها للحصول على درجة الماجستير في علم اللغة.

المراجع العربية

◇ **بحيري (سعيد حسن)**: نظرية التبعية في التحليل النحوي،مكتبة الأنجلو المصرية، ١٩٨٨م.

◇ **حسان (تمام)**: مناهج البحث في اللغة، مكتبة الأنجلو القاهرة، ط١، ١٩٥٥م.

◇ **الدحداح (أنطوان)**: معجم لغة النحو العربي، مكتبة لبنان ناشرون، ٢٠٠م.

◇ **روبي (أحمد)**: بناء بنك شجري نحوي للغة العربية الفصحى المعاصرة (لغة الصحافة الإلكترونية المصرية نموذجًا)، رسالة ماجستير، كلية دار العلوم-جامعة الفيوم، ٢٠١٦م.

◇ **زكريا (ميشال)**: الألسنية التوليدية وقواعد اللغة العربية (النظرية الألسنية)، ط٢، المؤسسة الجامعية للدراسات والنشر والتوزيع، بيروت،١٩٨٦م.

◇ **السعيد (المعتز بالله)**: مدونة معجم عربي معاصر «معالجة لغوية حاسوبية» رسالة ماجستير، جامعة القاهرة، ٢٠٠٨م.

◇ **السعيد (المعتز بالله)**: مدونة معجم تاريخي للغة العربية «معالجة لغوية حاسوبية»، رسالة دكتوراة، جامعة القاهرة، ٢٠١١م.

◇ **شريف (عمرو)**: ثم صار المخ عقلا، طبعة مكتبة الشروق الدولية، ط ٢، ٢٠١٣م.

◇ **شمس الدين (جلال)**: الأنماط الشكلية لكلام العرب، نظرية وتطبيقا دراسة بنيوية، مؤسسة الثقافة الجامعية، الأسكندرية، ط١، ١٩٩٥م.

◇ **عاشور (المنصف)**: بنية الجملة العربية بين التحليل والنظرية، منشورات كلية الآداب بمنوبة، ١٩٩١م.

◇ **عفيفي (أحمد مصطفى)**:النظرية النحوية-المفاهيم والتحديات، وقائع مؤتمر«العربية وقرن من الدرس النحوي «دار العلوم-القاهرة» ٢٠٠٣م.

◇ **علي (نبيل)، حجازي (نادية)**: الفجوة الرقمية «رؤية عربية لمجتمع المعرفة»،عالم المعرفة،٢٠٠٥م.

◇ **علي (نبيل)**: اللغة العربية والحاسوب، تعريب، ١٩٨٨م.

◇ **غاليم (محمد)**: هندسة التوازي النحوي وبنية الذهن المعرفية. كتاب آفاق

اللسانيات (تكريما للأستاذ الدكتور نهاد الموسى)، مركز دراسات الوحدة العربية، بيروت، ط ١، ٢٠١١م.

◇ **الفهري، (عبدالقادر الفاسي)**: البناء الموازي: نظرية في بناء الكلمة وبناء الجملة، دار توبقال للنشر، الدار البيضاء، ١٩٩٠م.

◇ **الموسى، (نهاد)**: العربية، نحو توصيف جديد في ضوء اللسانيات الحاسوبية، المؤسسة العربية للدراسات والنشر، بيروت، ٢٠٠٠م.

المراجع الأجنبية

◇ **Abeillé, A. & Rambo, O.**(2000)Tree Adjoining Grammar Formalisms, Linguistic, Analysis and Processing ,center for the study of languge and in formation.

◇ **Abeillé, A**. (2003). Treebanks: Building and Using Parsed Corpora. Springer Science & Business Media.

◇ **Attia, M**. (2007). Arabic Tokenization Systems. In proceeding of ACL.

◇ **Attiya, M**. (2004).Theory and Implementation Of a Large-Scale Arabic Phonetic Transcriptor, and Applications. PhD dissertation, Faculty of Engineering, Cairo University, P. 42.

◇ Dependency Viewer. [Version 1] [Computer Software] **Tang, G**: Nanjing university.

◇ **Dirven, R. & Langacker, R**. (1992). Grammar in Mind and Brain. Mouton de Gruyter, Berlin. New York.

◇ **Fraj, F. & Othmane, ch. & Ahmed, M.** (2008). ArabTAG: A tree adjoining grammar for Arabic syntactic structures. ACIT 2009, Tunisia, Hammamet.

◇ **Garsid, R., Leech, G. & McEnery, T**. (2013). Corpus Annotation. Second Published by Routledge. New York USA.

◇ **Gibbon, D. & Mertins, I. & Moore, R** .(2000).Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation. Springer Science & Business Media.

◇ **Habash, N. & Faraj, R**. (2009). Syntactic Annotation in the Columbia Arabic Treebank. In Proceeding of ELDA.

◇ **Hale, J. & Callaway, F.** (2014). Modeling Neural Correlates of Syntactic Structure Building. In AMLaP 2014-Poster, Scotland.

◇ **Jurafsky.D, Martin.J.** (2007). Speech and Language Processing: An introduction to natural language processing. Second Edition.

◇ **Kennedy, G.** (1998). An Introduction to Corpus Linguistics. Longman. P.223.

◇ **Pustejovsky, J. & Stubbs, A**. (2012). Natural Language Annotation For Machine Learning. Frist Edition.O'Reilly Media.

◇ **Rambow, O**. (2010). The Simple Truth about Dependency and Phrase Structure Representations.Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, Los Angeles, California.

◇ **Schmidt, P.**(2005). Mildly Context Sensitive Grammar Formalisms.

◇ **Sharaf, A. & Atwell, ES**. (2010). Arabic and Quranic computational linguistics projects at the University of Leeds. In Proceedings of the workshop of Increasing Arabic Contents on the Web, organized by Arab League Educational, Cultural and Scientific Organization (ALECSO).

◇ **Van Halteren, H.** (1999). Syntactic Wordclass Tagging. Springer-Science+Business Media.B.V. University of Nijmegen.

◇ ◇ ◇

الفصل الرابع

التحليل الدلالي

د. إشراق علي أحمد الرفاعي[1]

ملخص البحث

يتطرق الفصل إلى علم التحليل الدلالي باعتباره فرعاً من فروع اللسانيات الحاسوبية ومعالجة اللغات الطبيعية، يقدم الفصل نبذة تعريفية عن التحليل الدلالي تشمل استعراضا لأهم المصطلحات المرتبطة بهذا المجال مثل المتضادات والمترادفات، إضافة للفرق بين المعنى الحرفي والمعنى العملي للنصوص، كما يشير الفصل إلى المنهج البحثي المستخدم في دراسة التحليل الدلالي، إضافة إلى أبرز الموارد اللغوية المتاحة مثل شبكة الكلمات العربية (Arabic WordNet). يتطرق الفصل فيما يلي ذلك إلى عدد من أبرز تطبيقات هذا المجال وهي: تحليل العواطف، وفك اللبس الدلالي، مع تعريف

١- أستاذ مساعد في كلية علوم الحاسب الآلي بجامعة جازان. درست الماجستير والدكتوراه في جامعة هيروت وات الأسكتلندية. لها عدة أبحاث منشورة في اللسانيات الحاسوبية وتطبيقاتها لمعالجة نصوص اللغة العربية والإنجليزية والتي تتمحور حول تحليل المشاعر باستخدام خوارزميات التعلم الآلي، إضافة إلى دراسة معالجة نصوص الفصحى واللهجات المحلية، عملت الكاتبة كمحكمة لدى عدد من الدوريات العلمية والمؤتمرات الدولية. أنشأت عدداً من المدونات اللغوية التي خدمت عدداً كبيراً من الباحثين المهتمين بمجال تحليل العواطف في اللغة العربية والتي تم الرجوع لها أكثر من ٢٠٠ مرة (حسب موقع ELRA)، كما شاركت في مسابقات عالمية للسانيات الحاسوبية مثل سلسلة SemEval والتي تعد أكبر مسابقة للتحليل الدلالي لأهم لغات العالم مثل العربية والإنجليزية، وقد حصدت المركز الأول في فرع اللغة العربية في العام ٢٠١٦م. (eshragrefaee@gmail.com)

كل منهما، واستعراض أبرز ما أنجز فيهما من أبحاث وتطبيقات، كما يتطرق الفصل إلى الحديث عن الكينونات كمفهوم مهم عند دراسة التحليل الدلالي، ويُعنى بدراسة الكلمات المجردة والعلاقات فيما بينها من حيث المعنى. في الجزء الأخير من الفصل نشير إلى أبرز الجهود البحثية في مجال التحليل الدلالي فيما يخص اللغة العربية، والتي قُدمت من قبل مجموعات بحثية شهيرة حول العالم؛ حتى يتسنى للقارئ الاطلاع على المخرجات البحثية الأحدث في هذا المجال.

◇◇◇

الفصل الرابع: التحليل الدلالي

١. مقدمة

التحليل الدلالي هو أحد أبرز وأهم فروع معالجة اللغة الطبيعية (Natural Language Processing -NLP)، فهو تخصص يُعنى بدراسة معنى التعابير اللغوية (Linguistic Expression) (حبش والخليفة، ٢٠١٤)، وله العديد من التطبيقات الحياتية الهامة التي سوف نستعرضها خلال هذا الفصل.

٢. تعريف

في علم اللغويات يُعرف التحليل الدلالي على أنه دراسة معنى الكلمات في السياق، ويشير مانينق وشوتزي (Manning and Schütze, 1999) إلى إمكانية تقسيم دراسة الدلالات (Semantics) إلى شقين:

١. دراسة معنى الكلمات منفردة (Individual Words).
٢. دراسة المعنى الإجمالي للكلمات مجتمعة في عبارات أو جمل (Combined Meaning).

في الشق الأول تظهر لدينا مفاهيم أساسية، أبرزها:

المشترك اللفظي أو الجناس (Homonyms) والتي يعرفها حبش والخليفة (٢٠١٤) على أنها حالة وجود كلمتين متشابهتين في الإملاء والنطق ولكنهما مختلفتين في المعنى،

على سبيل المثال كلمة «بيت» قد تعني مكان العيش أو بيت الشِعر.

الجناس الخطي (Homograph) وهي الحالة التي يكون للكلمات فيها نفس الإملاء ولكن النطق يختلف، على سبيل المثال كلمة «حب» دون تشكيل قد تُنطق بضم الحاء «حُب» أو فتحه «حَب»، وفي كل من الحالتين سيكون المعنى مختلفا لهذه الكلمة.

التطابق اللفظي (Homophony) هي أن يكون للكلمات نفس النطق ولكن مع الرسم الإملائي يختلف، على سبيل المثال كلمتي عصى وعصا.

المتضادات (Antonyms) هي كلمات تحمل معاني متضادة، على سبيل المثال بارد وحار طويل وقصير.

المترادفات (Synonyms) وهي كلمات مختلفة ولكنها تحمل نفس المعنى، مثل سيارة ومركبة وكذلك بيت ودار.

٣. التحليل الدلالي في اللسانيات الحاسوبية

انتهج علماء اللغة بناء مدونات لغوية ضخمة تحوي مفردات اللغة وعلاقتها ببعضها من حيث المعنى (كونها متضادات أو مترادفات على سبيل المثال)، ومن أشهر تلك المدونات شبكة الكلمات (WordNet) (Fellbaum, 1998)، وهي قاعدة بيانات معجمية قابلة للقراءة بواسطة أجهزة الحاسب وتضم عدداً كبيراً جداً من الكلمات على شكل مجموعات يطلق عليها المجموعات المترادفة (Synsets)، إذ تُعتبر كل مجموعه مترادفة ممثلةً لمفهوم فريد من نوعه وتضم كل المرادفات المعرفية (Cognitive synonyms) لذلك المفهوم، فمثلا في نفس المجموعة المترادفة التي تضم كلمة "بارد" قد نجد كلمات مثل "شتاء" و "قارس" و"تجمد" ومن الممكن أن ترتبط كلمة "يتبع" بمفردات مثل "يلحق" و"يتعقب" و"يطيع" (Jaap and Marx, 2002)، وترتبط كل مجموعة مترادفة (Synset) بدورها مع مجموعات أخرى باستخدام العلاقات الدلالية مثل علاقات التضمين (Hyponomy)، فمثلا كلمة "رجل" تندرج تحت مجال أوسع وهو كلمة "إنسان" وكلمة "قطار" تندرج تحت مجال "وسائل النقل"، وكذلك علاقة الاشتمال (Hypernymy) وهي علاقة اشتمال كلمة في مجال دلالي لكلمة أخرى، فمثلا كلمة "حيوان" شاملة لكلمة "قطة" (Fellbaum,1998)، انظر الرسم التوضيحي في الشكل ١.



الشكل ١: مثال لارتباط الكلمات داخل شبكة الكلمات بناء على معانيها (Simpson and Dao, 2016).

وتعد شبكة بريستون الإنجليزية أول شبكة من نوعها في هذا المجال (Fellbaum, 1998; Princeton University, 2010)، وفي العام ٢٠٠٦ بدأت الجهود بالتضافر لبناء شبكة مماثلة للغة العربية (AWN) – Arabic WordNet والتي تستند في تصميمها على شبكة بريستون الإنجليزية، إذْ رُبطت مجموعة المترادفات العربية بما يقابلها في الشبكة الإنجليزية من حيث المعنى.

بعد أن يتم تحديد معنى الكلمات منفردة عبر وسائط مثل شبكة الكلمات فعندئذ يتم التوجه إلى تجميعها لتحديد المعنى الإجمالي للجملة بناء على تقسيم مانينق وشوتزي (Manning and Schütze, 1999) الذي سبقت الإشارة إليه.

٣,١ المعنى الحرفي أم المعنى الفعلي Semantics vs. Pragmatics؟

في تحديد المعنى الإجمالي للكلام يظهر لدينا مفهومان مهمان هما:

١. **المعنى الحرفي (Semantics)** وهو معنى الكلمات في اللغة بناء على موقعها من الإعراب، مثلا كلمة «هم» يتغير معناها حسب موقعها الإعرابي، فإما أن تكون ضميرا «هُم يلعبون بالكرة» أو اسماً كما في «أدركني في هذا الشأن هَم وحزن».

٢. **المعنى الفعلي / العملي (Pragmatic)** وهو المعنى الذي نستنتجه بناء على معرفتنا لسياق الكلام (Crass, 2012)، فمثلا "حضر موت" معناها الحرفي هو حضور الموت ولكن غالبا ما سيكون معناها الفعلي هو مدينة حضرموت اليمنية المعروفة، مثال آخر

هو كلمة "عين" معناها المباشر غالبا هو العين البشرية، ولكن معناها الفعلي قد يتغير حسب السياق الذي تَرِد فيه لتعني:

حرف العين، أو عين الماء، أو عين الحسد، أو مدينة العين في دولة الإمارات.

ما تقدم له دلالة هامة، وهي أن تحديد المعنى الكلي للجملة مشكلة معقدة (Complex problem)، مما جعلها مسألة ذات أهمية بالغة في معالجة اللغة الطبيعية واللسانيات الحاسوبية، إذْ إن اللغة الطبيعية لا تخضع دائما لمبدأ تراكبية المعنى(Principle of compositionality وهي أن معنى الجملة الكلي يمكن معرفته بالضرورة بناء على المعنى الفردي للكلمات المكونة لتلك الجملة، (Manning and Schütze, 1999) فكما اتضح لدينا من الأمثلة السابقة يمكن لمعنى الكلمات أن يتغير بتغير موقعها الإعرابي وكذلك بتغير سياق الكلام.

٣, ٢ التعبير المجازي (Idioms)

أحد الجوانب الهامة التي يجدر الإشارة إليها عند دراسة الدلالات(Semantics) هو التعبير المجازي، وهو استخدام الكلمات في غير معناها الظاهر، فمثلا قولنا "رأيت أسدا يكر على العدو بسيفه" يتضح منها أن الأسد هو استخدام مجازي يقصد به الشخص صاحب الشجاعة والإقدام وليس المعنى الظاهر وهو أن الأسد يحمل سيفا.

التحليل الدلالي له عدة تطبيقات هامة في مجال اللسانيات الحاسوبية، أبرزها هو فك اللبس الدلالي (Word sense disambiguation) وتحليل المشاعر (Sentiment analysis) والتي سنتطرق إليها في ثنايا هذا الفصل.

٤ . فك اللبس الدلالي Word Sense Disambiguation

يُعَرِّف الموجي وآخرون (Elmougy et al., 2008) فك اللبس الدلالي (WSD) بأنه عملية اختيار معنى لكلمة تحمل معاني متعددة بحيث يتناسب ذلك المعنى مع السياق الذي تظهر فيه تلك الكلمة وبحيث يكون الاختيار من مجموعة معانٍ معروفة ومحددة مسبقا.

"Word Sense Disambiguation (WSD) is the process of selecting a sense of an ambiguous word in a given context from a set of predefined senses".

ما تم التطرق له في الفقرات السابقة يبرز أهمية فك البس الدلالي كأحد أهم تطبيقات معالجة اللغة الطبيعية واللسانيات الحاسوبية في مجال دراسة التحليل الدلالي، والذي يُعنى بتحليل معنى الكلمات وتحديد المعنى الإجمالي لكلمات أو عبارات أو جمل في سياقها المعطى، بعبارة أخرى، إن بعض الكلمات يمكنها أن تحمل أكثر من معنى(senses) ، ويمكن من خلال توظيف السياق المحيط بها تحديد معناها المقصود في السياق بدقة أكبر (Manning and Schütze, 1999)، فكلمة "عين" والتي تحمل أكثر من معنى(senses) كما تقدم الإشارة إلى ذلك يمكن أن نعرف معناها بدقة أكبر إذا علمنا أنها استُخدمت في سياق "مدينة العين".

إن الدراسات والأبحاث في مجال اللسانيات الحاسوبية والتي تطرقت لمسألة فك اللبس الدلالي استخدَمت فرضية مفادها أن كل كلمة لها عدد محدد من المعاني (senses) المختلفة، والتي يمكن تخزينها في قاموس يضم الكلمات ومعانيها أو أي مخزن لغوي، بعد ذلك يُستخدم برنامج حاسوبي للبحث عن المعاني المختلفة لأي كلمة معطاة داخل مخازن الذخيرة لاستعادتها، ثم يقوم بعملية اتخاذ القرار لتحديد أي معنى هو الأقرب للصواب في سياق الكلام المعطى، هذه البرامج الحاسوبية غالبا ما تعتمد على خوارزميات التعلم الآلي (Machine Learning Algorithms).

النمط البحثي السابق في فك اللبس الدلالي غالبا ما يُستخدم لتحديد معاني الكلمات التي تحمل نفس الصفة النحوية (مثلا: كلاهما اسم أو كلاهما فعل)، لكن ماذا عن الحالات التي تختلف فيها الصفة النحوية للكلمة؟ مثلا "هُم" و "هَم" إذ إن الأولى ضمير منفصل والأخرى اسم، يعتبر وسم أجزاء الكلام (Part of Speech Tagging – POS) هو أحد أشكال فك اللبس الدلالي الذي يمكن استخدامه في مثل هذه الحالات، كما أشار إلى ذلك مانينق وشويتز (Manning and Schütze, 1999)[1].

١- ينصح بالنظر هنا للفصل الثاني من كتاب مانينق وشويتز (Manning and Schutze, 1999) والذي يتطرق لمزيد من التفصيل للتحليل الصرفي في مجال اللسانيات الحاسوبية وكذلك ينصح بالرجوع لكتاب حبش والخليفة (٢٠١٠) لمعرفة أدوات التحليل الصرفي المتاحة للغة العربية.

إن زيادة الدقة في تحديد المعنى الصحيح لكلمة ما في سياقها المعطى أمر بالغ الأهمية، خصوصا وأن هناك عددا من تطبيقات اللسانيات الحاسوبية يمكن أن تعتمد بشكل كبير على دقة نتائج فك اللبس الدلالي، على سبيل المثال، الترجمة الآلية (Machine Translation)، فعند ترجمة كلمة "عين" باعتبارها مدينة ستظهر كـ "City of Ain" بينما تُترجم إلى "eye" عند الإشارة إلى العين البشرية.

٤ , ١ الموارد اللغوية اللازمة في أنظمة فك اللبس الدلالي (Resource Requirement)

التطرق إلى مسالة فك اللبس الدلالي قد يتفاوت بناءً على الموارد والذخيرة اللغوية المتاحة لبناء نظام حاسوبي يقوم بفك اللبس الدلالي تلقائيا، حيث إن الخوارزميات المستخدمة غالبا ما تحتاج إلى مدونة لغوية متاحة يتم تغذية الخوارزمية بها بهدف بناء وتدريب نموذج رياضي (Statistical model) يكون بعد ذلك قادرا على القيام بتحديد معنى الكلمات في سياقها آليا، فهناك نهج بحثي يعتمد على وجود عينة تدريبية (Training examples) غالبا ما تكون معدة يدويا (Supervised disambiguation)، وهناك بحوث أخرى قد تعتمد على استخدام قواميس لغوية ضخمة (Dictionary-based disambiguation)، ويتم تقييم أداء النموذج الرياضي والنظام الحاسوبي الذي يُبنى عليه باستخدام عينة اختبار (Testing examples) تحتوي على عدد من الكلمات التي تم فك لبسها يدويا، على سبيل المثال كلمة "عين" في مثال "العين هي من أجمل مدن دولة الإمارات" ستحمل التأشير التالي:

{(الكلمة (word): العين، الدلالة في السياق (sense label): مدينة العين}.

تَوفرُ عددٍ كبير من هذه الأمثلة (بالمئات أو بالآلاف) ضروري لبناء أنظمة حاسوبية مدربة على القيام بفك اللبس الدلالي آليا (Supervised systems) بدقة عالية، ولكن تجدر الإشارة إلى أن توفير مثل هذه الموارد مكلف جدا من ناحية الوقت والجهد اللازمين (Knowledge sources)، وهذا ما أدى بالباحثين للنظر في طرق بديلة لبناء الأنظمة الآلية، وذلك إما باستخدام القواميس اللغوية الموجودة والتي تحتوي على عدد كبير من الكلمات ومعانيها في أكثر من سياق (انظر المثال في الشكل ٢)، أو بتكوين عينة تدريبية ذات حجم بسيط نسبيا واستخدامها في بناء النظام مبدئيا، ثم السماح له بالتعلم

تدريجيا بتعريضه لأمثلة خام / غير معدة (Unlabeled examples) مع وجود مراقبة مستمرة من قبل مُطوري النظام للتعديل والتصحيح بشكل دائم، وهذه الطريقة تُعْرف بالمراقبة الجُزئية (Semi-supervised learning)[١].


معنى عين في معجم المعاني الجامع - معجم عربي عربي

1. عَيَّنَ: (اسم)
عَيَّن : مصدر عَيِنَ

2. عَيِنَ: (فعل)
عَيِنَ عَيَنًا ، وعِينَةً
عَيِنَ : عَيِنَ اتَّسعت عينُه وحسُنت
فهو أعينُ ، وهي عَيناء والجمع : عِينٌ

3. عَيْن: (اسم)
عَيْن : مصدر عانَ

4. عَيَّنَ: (فعل)
عيَّنَ يعيِّن ، تعيينًا ، فهو مُعيِّن ، والمفعول مُعيَّن
عَيَّنَ الرجلُ : أخذ أو أعطى بالعِينة : أي السَّلَف
عَيَّنَ التاجرُ : باع سِلْعَتَهُ بثمن إلى أجل ، ثم اشتراها من المشتري في المجلس نفسه بأقلَّ من ذلك الثمن نقدًا ؛ ليسلَمَ من الرِّبا
عَيَّنَ الشجرُ : نَضِرَ ونَوَّر

الشكل ٢: مثال يوضح نتائج البحث عن كلمة "عين" في معجم المعاني الجامع، المصدر موقع المعاني.

٤, ٢ فك اللبس الدلالي في اللغة العربية

توجد عدد من الأبحاث التي تطرقت لاستخدام اللسانيات الحاسوبية لفك اللبس الدلالي في اللغة العربية، فقد استعرض الموجي وآخرون (Elmougy et al., 2008) مجموعة تجارب في هذا المجال توصلوا فيها إلى أن التجذير أو إرجاع الكلمات إلى جذورها ساهم بشكل كبير في رفع دقة البرنامج الآلي الذي طوره الباحثون لفك اللبس الدلالي في اللغة العربية، والذي اعتمد على واحدة من أبرز خوارزميات التعلم الآلي وهي Naïve Bayes ، كما توصل الباحثون إلى أن استخدام هذه الطريقة ساهم بشكل فعال في تخفيف اللبس الناتج من عدم وجود التشكيل في معظم النصوص العربية، إذ يؤثر عدم وجود التشكيل في خلق مساحة كبيرة من اللبس والغموض في معاني الكلمات، كما أشرنا في مثال سابق في مسألة التفاوت في معنى "هَمْ" و "هُمْ".

١- لمزيد من المعلومات حول خوارزميات التعلم الآلي المستخدمة والفرق بينها في الأداء ينصح بالرجوع إلى الفصل السابع من كتاب مانينق وشويتز (Manning and Schutze, 1999).

في دراسة حديثة، قدمت ناديا بوحريز وآخرون (Bouhriz et al., 2016) مجموعة من التجارب، وتوصلوا إلى أنه بالإضافة إلى الاعتماد على السياق داخل الجملة / السياق المحلي (Local context) لفك لبس معنى كلمة معينة كما هي الحال في جُل أبحاث فك اللبس الدلالي، يمكن كذلك الاستفادة من السياق في الجمل السابقة واللاحقة / السياق العام (Global context) محققا معدل دقة قدره ٧٤٪ عند تجربته على نصوص عربية مأخوذة من مصادر إخبارية.

على الرغم من التقدم الذي أحرزته أبحاث اللبس الدلالي عند تطبيقها على اللغة العربية إلا أن مزيدا من الأبحاث المستقبلية والعمل على إنشاء المزيد من الذخيرة اللازمة لبناء وتدريب الأنظمة الآلية قد يكون له دور فعال في رفع مستوى دقة الأداء.

٥. تحليل المشاعر (Sentiment Analysis)

أحد اتجاهات البحث الحديثة ضمن التحليل الدلالي هي التوجه لتحليل المشاعر وتوجهات الرأي (Sentiment Analysis)، وهو مجال يُعْنَى بدراسة وتحليل قطبية المشاعر في نص ما، بمعنى تحديد اتجاه المشاعر المعبر عنها، بحيث تكون إما إيجابية أو سلبية أو محايدة، ويعرف ليو (Liu, 2012) هذا العلم على أنه علم لتصنيف النص بحسب المشاعر التي يحتويها إلى إيجابي أو سلبي أو محايد، آخذا بعين الاعتبار وجهة نظر كاتب النص وليس وجهة نظر قارئه (انظر الأمثلة في الجدول التالي):

مثال	قطبية المشاعر
تنظيم رائع ومتميز في قمة دبي الحكومية لهذا العام.	ايجابي
تنحى الدكتاتور مبارك عن سدة الحكم.	سلبي
يوجد آيفون بين كل أربعة أجهزة ذكية.	محايد

جدول ١: أمثلة لنصوص ذات قطبيات مختلفة، إيجابية وسلبية ومحايدة.

ويفرق ليو (Liu, 2012) بين تحديد اتجاه قطبية المشاعر المعبر عنه: فهي إما أن تكون من وجهة نظر كاتب النص أو من وجهة نظر قارئه، فمثلا قراءة خبر عن توسع المستوطنات الإسرائيلية في غزة غالبا ما سيكون خبرا سلبيا للقارئ الفلسطيني، وفي

نفس الوقت سيكون محايدا بل ربما إيجابياً لشخص على الطرف الآخر، وعلية ارتأت معظم الأبحاث في تحليل المشاعر أن تعمل على تحديد قطبيتها من وجهة نظر كاتب النص (Author perspective) وليس من وجهة نظر قارئه (Reader perspective).

إن التطبيقات العملية لتحليل المشاعر كأحد تفرعات معالجة اللغة الطبيعية واللسانيات الحاسوبية متعددة وذات تأثيرات لها أبعاد مختلفة، فمن وجهة نظر اللسانيات الحاسوبية يُنظر إلى تحليل المشاعر بأنه أحد تطبيقات تصنيف النصوص (Text classification problem) التي حققت تقدماً كبيراً عند تطبيقها على اللغة العربية (الفصحى المعاصرة Modern Standard Arabic –MSA) ووصلت إلى معدلات دقة عالية، فشانتر (Chanter, 2013) مثلاً وصل إلى دقة تجاوزت ٩٥٪ عند تصنيف نصوص النشرات الإخبارية العربية، إذ تُحدَّدُ فئة النص تلقائيا إلى: أخبار رياضية، اقتصادية، وهكذا، وقد استخدم الكاتب إحدى خوارزميات التعلم الآلي (Machine learning algorithm) والتي قام بتطويرها من خلال أبحاثه لتكون أكثر دقة وسرعة، ولكن بالنسبة لتحليل المشاعر (كونه مسالة تصنيف نصوص كذلك كما سبق وأشرنا) أظهرت الأبحاث انخفاضا كبيرا في الأداء ليكون ٦٠-٧٠٪ في اللغة الإنجليزية (Nakov et al., 2016) و٥٢-٦٥٪ في اللغة العربية(Abdul-Mageed 2015 ,)، وجدير بالذكر هنا أن أبحاث تحليل المشاعر ابتعدت عن النصوص التقليدية (مثل النصوص الإخبارية) وتوجهت إلى منصات شبكات التواصل الاجتماعي (مثل تويتر وفيسبوك)، وذلك بحثا عن مصدر غني بالتعبيرات ذات الدلالة العاطفية، حيث برزت شبكات التواصل الاجتماعي في السنوات الأخيرة لتكون حيزا يسمح لعدد هائل من المستخدمين يمثلون فئات عمرية مختلفة وخلفيات دينية وثقافية متعددة بالتعبير عن آرائهم وتوجهاتهم ومشاعرهم تجاه مواضيع وأشخاص أو حتى منتجات تجارية مختلفة، وهو ما جعل تحليل المشاعر - باستخدام تقنيات اللسانيات الحاسوبية وأدوات معالجة اللغة الطبيعية - للنصوص التي يتم ضخها بكميات كبيرة وبشكل يومي عبر شبكات التواصل الاجتماعي ذات أهمية بالغة في تطبيقات حياتية متعددة، منها:

- تقييم مدى نجاح منتج أو خدمة تم إصدارها مؤخرا، كتقييم شعور المستخدمين حول أحدث إصدار لأحد الهواتف الذكية، وقد أشار ليو (Liu, 2012) إلى أن الشركات العملاقة مثل قوقل وميكروسوفت لديها أنظمة حاسوبية لتحليل

المشاعر تم تصميمها وبناؤها بشكل يخدم أهداف تلك الشركات.

- تحديد مدى شعبية أحد المرشحين السياسيين أو الأحزاب السياسية، إذ إن منصات شبكات التواصل الاجتماعي تكون نشطة خلال فترات الانتخابات السياسية، وتزخر بالكثير من الآراء (المؤيدة أو المعارضة) تجاه السياسيين أو الأحزاب السياسية إبان فترة الانتخابات (Pang and Lee, 2008).
- التنبؤ بأداء أسواق المال، حيث إن الأبحاث أثبتت ارتباطا بين أداء أسواق المال العالمية والأحداث التي تُسجل حول العالم، والتي تؤثر في مشاعر الناس وتنعكس على ما يعبرون عنه في شبكات التواصل الإلكترونية (Johan et al., 2011).
- أبحاث أخرى استخدمت تحليل المشاعر في نصوص شبكات التواصل لتقييم المزاج العام وقياس سعادة الشعوب (Public mood / National happiness) (Johan et al., 2011).
- كشف نزعات عنصرية أو آراء متطرفة، حيث قدم عباسي وآخرون (Abbasi et al., 2008) دراسة موسعة لاستخدام تحليل المشاعر في الكشف عن توجهات عنصرية في مواقع اجتماعية عربية وإنجليزية.

٥, ١ مميزات وتحديات تحليل المشاعر و شبكات التواصل الاجتماعي؟

تتميز نصوص شبكات التواصل الاجتماعي بكونها مصدرا غنيا للنصوص التي يمكن توظيفها في تطبيقات اللسانيات الحاسوبية وعلى رأسها تحليل المشاعر، كونها تزود الباحثين بكمية كبيرة من النصوص التي يمكن جمعها بشكل مجاني، حيث إن جزءاً كبير من النصوص التي تُستخدم في أبحاث معالجة اللغة الطبيعية يتطلب الوصول إليها شراء حقوق الاستخدام لهذه النصوص، وأبرز المزودين لهذه الخدمة هي جمعية البيانات اللغوية التي تديرها جامعة بنسلفانيا في الولايات المتحدة (Linguistic Data Consortium-LDC) وتضم مدونات لغوية كبيرة لعدد من اللغات ومنها اللغة العربية.

كذلك تتميز شبكات التواصل بتأثيرها الواسع، إذ إن الآراء المطروحة خلالها تبلغ شريحة واسعة من المجتمع، وفي السنوات الأخيرة سببت الآراء التي يتم بثها عبر منصات التواصل الاجتماعي تأثيرات اجتماعية وسياسية ضخمة، منها على سبيل المثال الثورات والتغييرات السياسية والاجتماعية التي قامت في عدد من الدول العربية مؤخرا، والتي

كانت بدايتها حملات تأسست عبر تويتر وفيسبوك لتعبر عن آراء ومشاعر الأفراد تجاه الأنظمة السياسية في مجتمعاتهم (Buettner and Buettner, 2016) .

أبرز تحديات شبكات التواصل للنصوص العربية هي استخدام اللهجات المحلية (Local dialects) أكثر من الفصحى (MSA)، حيث إن أدوات اللسانيات الحاسوبية التي تم تطويرها حتى وقت قريب تركز على الفصحى، والقليل جدا من الأبحاث تطرق إلى اللهجات المحلية، وقد بحثت الرفاعي (Refaee, 2016) في إمكانية استخدام أدوات ومدونة لغوية تم تصميمها للفصحى واستخدامها على نصوص شبكات التواصل (والتي تُمثُل مزيجاً من الفصحى والعاميات) وتوصلت إلى أن هناك جدوى من استخدام مثل تلك الأدوات على الرغم من انخفاض الأداء العام مقارنة بالتجارب على الفصحى فقط، من أمثلة هذه الأدوات: أدوات المعالجة المبدئية الآلية للنصوص (Pre-processing tools) ، وكذلك أدوات التحليل الصرفي (Morphological analysers).

تحديات أخرى لتحليل المشاعر (وهي عامة وليست مقتصرة على اللغة العربية) تتمثل في استخدام اللغة غير المباشرة في التعبير عن المشاعر، فيمكن في اللغات الطبيعية التعبير عن مشاعر معينة (إيجابية أو سلبية) دون استخدام كلمات مباشرة ذات دلالة عاطفية، ومثل هذه النصوص قد يكون من السهل تحديد قطبيتها العاطفية عند قراءتها من قبل البشر، ولكنها تمثل تحديا للأنظمة الحاسوبية المصممة لتحليل المشاعر بشكل تلقائي، حيث إنها تعتمد وبشكل كبير على الكلمات ذات الدلالة القطبية الواضحة والقوية مثل "ممتاز" للقطبية الإيجابية و "بشع" للقطبية السلبية، وتستخدمها كعناصر (Features) أساسية في بناء أنظمة تحليل المشاعر (انظر المثال).

- أصبحت مصر مثل الفيلم الأجنبي الغير مترجم، الكل يتفرج ويترجم على مزاجه.

صعوبة أخرى تكمن في استخدام المشاعر المختلطة (التعبير عن مشاعر إيجابية وسلبية في جملة واحدة)، فعند التصنيف اليدوي لمثل تلك الأمثلة يتم التعامل مع هذه الحالة بتغليب المشاعر الأقوى، ولكن بالنسبة للأنظمة الآلية لتحليل المشاعر تمثل النصوص ذات المشاعر المختلطة أكبر مصدر للخطأ الذي يتسبب في خفض دقة هذه الأنظمة (Abbasi et al., 2014)، ومن النصوص ذات المشاعر المختلطة:

- **لست مع الإخوان سياسيا، ولكنني معهم إنسانيا.**

- **السنة والشيعة كل طرف يحمل صورة نمطية عن الآخر فيها الكثير من الزيف والحق.**

- **المساواة في قمع الحريات الشخصية عدل.**

تحدٍ آخر لتحليل المشاعر هو استخدام اللغة الهزلية (Sarcasm/Irony)، وهو استخدام كلمات إيجابية للتعبير عن مشاعر سلبية أو العكس، المثال التالي يوضح استخدام الكلمة الإيجابية "جميل" بصورة هزلية:

- **جميل هذا الصمت من الدول العربية لما يحدث في غزة.**

تعليمُ الأنظمة الآلية إدراك بعض الأنماط الهزلية يتطلبُ تزويد تلك الأنظمة بعدد كبير من الأمثلة للتدريب عليها (Training examples) وهو ما قد يصعب توفيره، خصوصا مع عدم توفر مدونة لغوية لخدمة هذا الغرض حتى الآن، ولكنها منطقة جديرة بالبحث مستقبلا.

٦. الكينونات (Ontologies)

أحد المفاهيم التي تجدر الإشارة إليها عند الحديث عن التحليل الدلالي هي الكينونات (Ontologies)، وهي عبارة عن مجموعة من المفاهيم المجردة التي ترتبط ببعضها وتملك كل منها مجموعة من الخصائص، ويُعرِّف قاموس أكسفورد الكينونات على أنها:

«مجموعة من المفاهيم والفئات في موضوع أو مجال ما، والتي تمتلك خصائص أو ملامح تُعرِّفها وتُعرِّف العلاقات الداخلية التي تربط فيما بينها».

"A set of concepts and categories in a subject area or domain that shows their properties and the relations between them" (١)

ومن أمثلة الكينونات شبكة الكلمات (WordNet) التي سبق الإشارة إليها (فقرة التحليل الدلالي في اللسانيات الحاسوبية)، ومن الكينونات الأخرى الشهيرة التي تتميز

1- https://www.oxforddictionaries.com/

بكونها متعددة اللغات شبكة بابل BabelNet التي تشمل أكثر من ٢٠٠ لغة ومنها اللغة العربية، والتي أنشأتها جامعة إسبانيزا في روما الإيطالية بطريقة آلية مع الاستعانة بشبكة WordNet وعدد آخر من المصادر مثل موسوعة Wikipedia، وذلك باستخدام الترجمة الآلية بين اللغات (Statistical machine translation)، وتعد شبكة بابل من أكبر المصادر اللغوية متعددة اللغات المتوفرة على الإطلاق.

تستخدم الكينونات في عدة مجالات يبرز من بينها الارتباط الدلالي للكلمات ومعانيها (Semantic relatedness)، وفك اللبس الدلالي عن طريق إيراد كل المعاني الممكنة لمصطلح أو كلمة معينة، حتى يتم توظيف خوارزميات اللسانيات الحاسوبية لاتخاذ قرار فيما يخص اختيار أدق المعاني للكلمة في سياق معين.

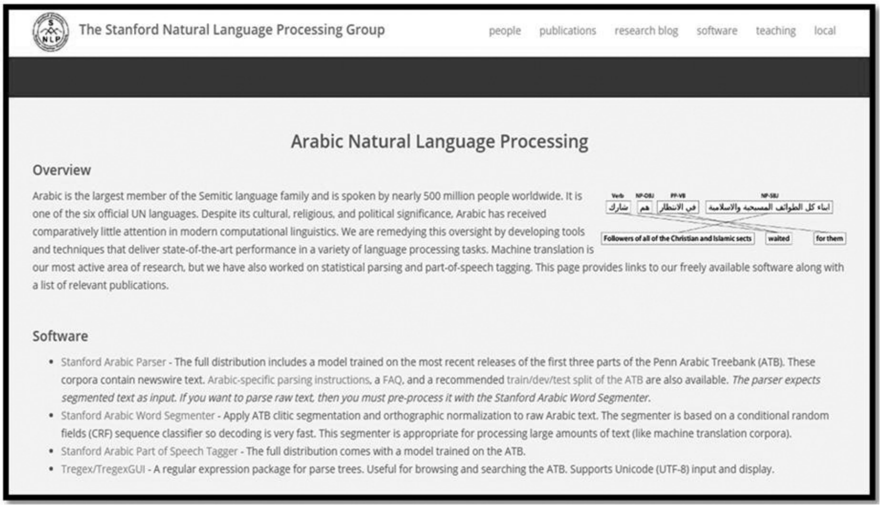


الشكل ٣: مثال يوضح مرادفات لكلمة "كتاب" مع ترجمتها لعدة لغات عبر شبكة بابل، المصدر BabelNet

٧. جهود بارزة في التحليل الدلالي للغة العربية

تجدر الإشارة إلى أبرز الجهود البحثية التي قُدِّمت ولاتزال فاعلة في مجال التحليل الدلالي وتطبيقاته (مثل: فك اللبس الدلالي وتحليل المشاعر) فيما يتعلق بمعالجة اللغة الطبيعية واللسانيات الحاسوبية، والتي تقدمها مجموعات بحثية هامة حول العالم أبرزها مجموعة the Stanford NLP group البحثية في جامعة ستانفورد بقيادة البروفيسور كريستوفر مانينق (Manning et al., 2008)، فهذه المجموعة البحثية تقدم جهودا متميزة لخدمة اللغة العربية نتج عنها حزمة من البرامج الفعالة والمدونات اللغوية المتميزة،

والتي قُدمت في جُلها بشكل مجاني لخدمة الباحثين في هذا المجال (انظر الصورة).



The Stanford Natural Language Processing Group   people   publications   research blog   software   teaching   local

Arabic Natural Language Processing

Overview

Arabic is the largest member of the Semitic language family and is spoken by nearly 500 million people worldwide. It is one of the six official UN languages. Despite its cultural, religious, and political significance, Arabic has received comparatively little attention in modern computational linguistics. We are remedying this oversight by developing tools and techniques that deliver state-of-the-art performance in a variety of language processing tasks. Machine translation is our most active area of research, but we have also worked on statistical parsing and part-of-speech tagging. This page provides links to our freely available software along with a list of relevant publications.

Software

- Stanford Arabic Parser - The full distribution includes a model trained on the most recent releases of the first three parts of the Penn Arabic Treebank (ATB). These corpora contain newswire text. Arabic-specific parsing instructions, a FAQ, and a recommended train/dev/test split of the ATB are also available. *The parser expects segmented text as input. If you want to parse raw text, then you must pre-process it with the Stanford Arabic Word Segmenter.*
- Stanford Arabic Word Segmenter - Apply ATB clitic segmentation and orthographic normalization to raw Arabic text. The segmenter is based on a conditional random fields (CRF) sequence classifier so decoding is very fast. This segmenter is appropriate for processing large amounts of text (like machine translation corpora).
- Stanford Arabic Part of Speech Tagger - The full distribution comes with a model trained on the ATB.
- Tregex/TregexGUI - A regular expression package for parse trees. Useful for browsing and searching the ATB. Supports Unicode (UTF-8) input and display.

الشكل٤ : الموقع الرسمي لمجموعة ستانفورد البحثية والمخصص لأبحاث اللغة العربية.

المراجع العربية

◇ **نزار حبش وهند الخليفة**: مقدمة في المعالجة الطبيعية للغة العربية، (الطبعة١)، دار جامعة الملك سعود للنشر، ٢٠١٤، صفحه ١٩٩-٢١٠.

◇ **معجم المعاني الجامع**: http://www.almaany.com/ar/dict/ar-ar//

المراجع الأجنبية

◇ **Fellbaum, Christiane**. WordNet. Blackwell Publishing Ltd, 1998.

◇ **Christiane Fellbaum**.WordNet: An Electronic Lexical Database. MIP Press, 1998, https://wordnet.princeton.edu/.

◇ **Manning, Christopher D. and Schütze Hinrich**. Foundations of statistical natural language processing. Vol. 999. Cambridge: MIT press, 1999.

◇ **Kamps, Jaap, and Maarten Marx**."Visualizing WordNet structure." Proc. of the 1st International Conference on Global WordNet. 2002.

◇ **Black, W., Elkateb, S., Rodriguez, H., Alkhalifa, M., Vossen, P., Pease, A., &Fellbaum, C**. Introducing the Arabic WordNet project. In Proceedings of the third international WordNet conference. 2006. (pp. 295-300).

◇ **Elmougy, Samir, H. Taher, and H. Noaman**. "Naïve Bayes classifier for **Arabic word sense disambiguation**." proceeding of the 6th International Conference on Informatics and Systems. 2008.

◇ **Abbasi, Ahmed, Hsinchun Chen, and Arab Salem**. "Sentiment analysis in multiple languages: Feature selection for opinion classification in Web forums." ACM Transactions on Information Systems (TOIS) 26.3 (2008): 12.

◇ **Pang, Bo, and Lillian Lee**. «Opinion mining and sentiment analysis.»Foundations and trends in information retrieval 2.1-2 (2008): 1-135.

◇ **Habash, Nizar Y.** "Introduction to Arabic natural language processing". Synthesis Lectures on Human Language Technologies 3.1 (2010): 1-187.

◇ **Princeton University**. “About WordNet.” WordNet. Princeton University. 2010. http://wordnet.princeton.edu.

◇ **Troy Simpson and Thanh Dao**. WordNet-based semantic similarity measurement. Code Project. http://www.codeproject.com/Articles/11835/WordNet-based-semantic-similarity-measurement.

◇ **Bollen Johan, Huina Mao, and Xiaojun Zeng**. “Twitter mood predicts the stock market.” Journal of Computational Science 2.1 (2011): 1-8.

◇ **Bollen Johan, Huina Mao, and Alberto Pepe**. “Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena.” ICWSM 11 (2011): 450-453.

◇ **Courtney Crass**. Pragmatic vs. Semantic Meaning: Sorting It Out. Bright Hub Education. 2012. http://www.brighthubeducation.com/english-homework-help/105856-understanding-pragmatic-vs-semantic-meaning/

◇ **Bing Liu**. Sentiment Analysis and Opinion Mining (Synthesis Lectures on Human Language Technology). Morgan & Claypool Publishers, 2012.

◇ **Christopher Manning, Dan Jurafsky and Percy Liang**. 2008. Arabic Natural Language Processing. The Stanford Natural Language Processing Group. http://nlp.stanford.edu/projects/arabic.shtml

◇ **Hamouda Khalifa Hamouda Chanter**. New Techniques for Arabic Document Classification. PhD thesis. The School of Mathematical and Computer Sciences, Heriot-Watt University. Edinburgh, United Kingdom. 2013.

◇ **Abbasi, Ahmed, Ammar Hassan, and Milan Dhar**. “Benchmarking Twitter Sentiment Analysis Tools.” LREC. 2014.

◇ **Muhammad Abdul-Mageed**. Subjectivity and Sentiment Analysis of Arabic as a morphologically-rich Language. PhD thesis. The School of Informatics and Computing, Indiana University, Bloomington, Indiana, United States. 2015.

◇ **Nadia Bouhriz, Faouzia Benabbon and El Habib Ben Lahmar**. Word Sense Disambiguation Approach for Arabic Text. In the International Journal of Advanced Computer Science and Application, vol. 7 No. 4. 2016. Pages 381-385.

◇ **Eshrag Refaee**. Sentiment Analysis for Micro-Blogging Platforms in Arabic. PhD thesis. The School of Mathematical and Computer Sciences, Heriot-Watt University. Edinburgh, United Kingdom. 2016.

◇ **Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., & Stoyanov**, V. "SemEval-2016 task 4: Sentiment analysis in Twitter."Proceedings of the 10th international workshop on semantic evaluation (SemEval 2016), San Diego, US. 2016.

◇ **Buettner, R and Buettner, K**. A Systematic literature review of Twitter research from a socio-political perspective. In the 49th Hawaii International Conference on System Science (HIVSS), pages 2206-2215. IEEE, 2016.

◇◇◇

الفصل الخامس

تحليل النصوص

د. صلاح راشد الناجم[1]

ملخص البحث

يتناول هذا المبحث أهمية تحليل النصوص كتطبيق أساسي من تطبيقات المعالجة الحاسوبية للغة الطبيعية (Natural Language Processing) وذلك من خلال اكتشاف وانتزاع معرفة هامة من نصوص حرة لا تسير وفق بنية منظمة (Unstructured Data). حيث أفرز التطور الكبير في مجال البيانات الضخمة (Big Data) كميات هائلة من البيانات النصية ومنها على سبيل المثال لا الحصر مشاركات وحوارات وسائل التواصل الاجتماعي. يتطلب تحليل هذه البيانات إيجاد تطبيقات ومنصات تحليلية ولغات برمجة وأدوات برمجية وخوارزميات (Algorithms) متخصصة للتعامل مع هذا الكم الهائل من البيانات النصية. وهنا تبرز أهمية تحليل النصوص كمجال بيني

(1) أستاذ مشارك في تخصص اللسانيات الحاسوبية والمعالجة الحاسوبية للغة الطبيعية بجامعة الكويت. حصل على شهادتي الماجستير والدكتوراه في اللسانيات الحاسوبية من جامعة أسيكس (Essex) في إنجلترا. نشر عددا من الأبحاث وشارك في تأليف كتب حول اللسانيات الحاسوبية والمعالجة الحاسوبية للغة العربية. كما قام بالإشراف على عدد من رسائل الماجستير والدكتوراه في جامعة الكويت وخارج دولة الكويت. كذلك شارك في تحكيم عدد من الأبحاث ومناقشة أطروحات الماجستير في جامعة الكويت. عمل كمستشار في عدد من الجهات الحكومية منها الديوان الأميري، مجلس الأمة، الجهاز المركزي لتكنولوجيا المعلومات، ووزارة الإعلام. مهتم باللسانيات الحاسوبية، المعالجة الحاسوبية للغة الطبيعية، تحليل النصوص، وتحليل البيانات الضخمة (Big Data). (salah.alnajem@ku.edu.kw)

(Interdisciplinary) إذ يدمج أكثر من مجال أكاديمي أهمها علم الحاسوب، اللسانيات الحاسوبية، استرجاع المعلومات (Information Retrieval)، تحليل البيانات (Data Mining)، تعلُّم الآلة (Machine Learning)، والإحصاء (Statistics). يدخل في هذا المجال انتزاع المعلومات (Information Extraction) من وثائق أو مواقع على الشبكة العنكبوتية أو غيرها، وتصنيف النصوص (Text Classification)، وانتزاع أسماء الكيانات (Named Entities)، وانتزاع العلاقات والأحداث، وتحليل المزاج العام (Sentiment Analysis). بعد التعريف بتحليل النصوص، يتحدث هذا المبحث عن أهمية البيانات الضخمة ومستويات ومراحل تحليل النصوص. ثم ينتقل إلى الحديث عن المعالجة الحاسوبية للنصوص وخطواتها، ثم يتناول أخيرا تطبيقات تحليل النصوص مثل تصنيف النصوص، انتزاع المعلومات، وتحليل المزاج العام.

◇◇◇

الفصل الخامس: تحليل النصوص

١ . تعريف

يعرف تحليل النصوص (Text Analysis /Text Mining) بأنه اكتشاف وانتزاع معرفة هامة من نصوص حرة، أي نصوص لا تسير وفق بنية منتظمة .(Unstructured Text) لتحقيق ذلك ، تُستَخدَم أنواع عديدة من التمثيل المعرفي (Knowledge Representation) للمعلومات اللغوية. نحصل على هذه المعلومات اللغوية عن طريق استخدام المعجم الإلكتروني (Lexicon) الذي يحوي كلمات لغة معينة وسماتها الصرف - النحوية وقيمها الافتراضية من حيث المزاج العام (Score Sentiment Default) إضافة إلى استخدام القوانين النحوية والمعلومات الدلالية كالتصنيف الدلالي (Ontology /Taxonomy) للكيانات والأحداث فضلا عن استخدام مكانز (Thesaurus) المترادفات والاختصارات.

يعد تحليل النصوص تطبيقا من تطبيقات المعالجة الحاسوبية للغة الطبيعية (Natural Language Processing). المعالجة الحاسوبية للغة الطبيعية هو فرع مشترك بين علم الحاسوب (Computer Science) واللسانيات الحاسوبية (Computational Linguistics). تعتبر المعالجة الحاسوبية للغة الطبيعية التطبيق العملي للجوانب النظرية لعلم الحاسوب و اللسانيات الحاسوبية .

يمثل تحليل النصوص مجالا بينيا (Interdisciplinary) حديثا يدمج أكثر من مجال أكاديمي أهمها علم الحاسوب، اللسانيات الحاسوبية، استرجاع المعلومات (Information Retrieval)، تحليل البيانات (Data Mining)، تعلُّم الآلة (Machine Learning)، والإحصاء (Statistics).

تُستخدم تقنيات تحليل النصوص في المجال التجاري والحكومي والأكاديمي وذلك لأن أغلب المعلومات الرقمية المستخدمة عالميا مخزنة على شكل نصوص لا تسير وفق بنية منظمة مقارنة مع البيانات التي تسير وفق بنية منظمة مثل البيانات الموجودة في قواعد البيانات (إيجناتو وميخاليكا ٢٠١٦، Ignatow & Mihalcea).

يدخل في هذا المجال انتزاع المعلومات (Information Extraction) من وثائق أو مواقع على الشبكة العنكبوتية أو غيرها وتصنيف النصوص (Text Classification) وانتزاع أسماء الكيانات (Named Entities) والعلاقات والأحداث وتحليل المزاج العام (Sentiment Analysis).

تتضمن عملية تحليل النصوص استخدام عدد من العمليات أهمها انتزاع المعلومات (Information Extraction) من مجموعة من النصوص وتطبيق مناهج إحصائية متقدمة واستخدام المعالجة الحاسوبية للغة الطبيعية من خلال التحليل الصرفي الآلي (Morphological Processing) والوسم الآلي للكلمات (Part-of-Speech Tagging) والتحليل النحوي (Syntactic Parsing) والتعرف على الكيانات (Entity Recognition) و استخدام معلومات معجمية وتقنيات إحصائية لمعرفة الكيانات في النصوص مثل أسماء الأشخاص والأماكن والشركات وغيرها. كذلك تشتمل هذه العمليات على ما يعرف بإزالة الغموض (Disambiguation) عن طريق استخدام معلومات سياقية لتحديد المعنى المقصود من الكلمة في حال وجود أكثر من معنى لها (ميخاليكا ٢٠٠٧، Mihalcea). كذلك تتضمن عملية تحليل النصوص تحليل المزاج العام (Sentiment Analysis) للنص والذي يهدف إلى تحديد معلومات متعلقة بالموقف والرأي كتحديد أن كلمة أو عبارة معينة في النص تحمل مدلولا إيجابيا أو سلبيا أو محايدا و تحديد العاطفة (emotion) المرتبطة بالكلمة أو العبارة (ستراباراڤا و ميخايكا ٢٠٠٨، Strapparava & Mihalcea).

٢. دور البيانات الضخمة

تسبب التطور في مجال البيانات الضخمة (Big Data) بإنتاج كميات هائلة من البيانات النصية. كذلك تسبب هذا التطور في إيجاد تطبيقات ومنصات تحليلية عديدة ولغات برمجة وأدوات برمجية وخوارزميات (Algotithms) متخصصة للتعامل مع هذا الكم الهائل من البيانات النصية. تعرف البيانات الضخمة بأنها مجموعات البيانات المركبة كبيرة الحجم والتي لا يمكن معالجتها باستخدام الوسائل اليدوية أو باستخدام تطبيقات معالجة البيانات التقليدية. من أمثلة البيانات الضخمة مشاركات وسائل التواصل الاجتماعي كالتغريدات ومشاركات المدونات (Blogs) وسجلات الشبكة العنكبوتية (Web Logs) وهي السجلات التي تنتجها أنظمة تحليل الشبكة العنكبوتية (Web Analytics) وترصد فيها سلوك زوار مواقع الشبكة وكيفية استخدامهم للمحتوى المنشور على هذه الصفحات مثل نظام Google Analytics. من أبرز التطبيقات التجارية المستخدمة في تحليل النصوص المشتقة من البيانات الضخمة نظام SAS Text Miner (انظر شاكرابورتي ٢٠١٤، Chakraborty). من لغات البرمجة المستخدمة في هذا المجال لغات Python و R، ومن الأدوات البرمجية مكتبات البرمجيات NLTK و GATE و Pandas و NumPy.

كما تسبب التطور في البيانات الضخمة في توافر مصادر متنوعة من البيانات النصية التي استُخدِمَت في أبحاث ومشاريع تحليل النصوص. من مصادر هذه البيانات المتعلقة بوسائل التواصل الاجتماعي أرشيف تويتر الرسمي (Twitte Gnip Firehose) والذي يوفر أرشيفا كاملا لتغريدات مستخدمي تويتر منذ إنشاء تطبيق تويتر إلى الآن.

في هذا السياق يقول المحلل السياسي جاري كينج، أن الجانب الثوري في البيانات الضخمة ليس حجم مجموعات البيانات (data sets) ولكن الجانب الثوري هو ما يستطيع الباحثون عمله الآن باستخدام هذه البيانات عن طريق الخوارزميات والأدوات البرمجية والتطبيقات المتخصصة في تحليل هذا النوع من البيانات، حيث أدى ذلك إلى الزيادة في استخدام التحليل الكمي في المجال الأكاديمي والعلمي والصناعي والحكومي (شاو ٢٠٠٤، Shaw).

وفي مجال متصل، يشير نيتين هاردينيا (هاردينيا ٢٠١٥، Hardeniya) إلى أن المهارات في مجال المعالجة الحاسوبية للغة الطبيعية تمثل إحدى أكثر المهارات ندرة

وهي مطلوبة بشكل كبير في مجال صناعة تقنية المعلومات. فبعد التطور الكبير في مجال البيانات الضخمة، صار التحدي الذي يواجه صناعة تقنية المعلومات هو إيجاد متخصصين يستطيعون التعامل ليس فقط مع البيانات التي تسير وفق بنية منظمة (Structured Data) كالمعلومات الموجودة في قواعد البيانات بل علينا إيجاد المتخصصين الذين يستطيعون التعامل مع البيانات التي تسير وفقا لبنية شبه منظمة (Semi-Structured) أو غير منظمة (Unstructured Data). في هذا السياق، نحن ننتج بيتابايتات (Petabytes) من البيانات على شكل تغريدات، مشاركات فيسبوك، مشاركات مدونات (Blogs) ، دردشات (Chats)، رسائل بريد إلكتروني، سجلات للشبكة العنكبوتية (Web Logs)، ومساهمات إبداء الرأي (Reviews). حيث تقوم الشركات بجمع هذه الأنواع المختلفة من البيانات لكي تتمكن من استهداف الشرائح المناسبة بشكل أفضل ولكي تحصل على استنتاجات ذات معنى من تحليلها. ومن أجل معالجة كل مصادر هذه البيانات التي تسير وفق بنية غير منظمة، يتطلب الأمر متخصصين في مجال المعالجة الحاسوبية للغة الطبيعية (هاردينيا ٢٠١٥، Hardeniya).

من جهة أخرى انتبهت الحكومات إلى أهمية التعامل مع البيانات الضخمة، حيث أدركت أن الحوار الذي يدور على وسائل التواصل الاجتماعي يمثل وسيلة حية لاستطلاع رأي الجمهور ولمعرفة اتجاه الرأي العام أو اتجاه فئة معينة في المجتمع مثل الشباب. كما يمكن من خلال هذا الحوار معرفة ردود أفعال الجمهور تجاه القضايا السياسية والاجتماعية والاقتصادية. كذلك يعد الحوار الذي يدور على الشبكات الاجتماعية والأنشطة التي ترتبط بها من المؤشرات الأساسية لقياس الأداء (Key Performance Indicators) والتي يستخدمها متخذو القرار والجهات الحكومية للتأكد من تحقيق الأهداف الاستراتيجية لاستراتيجياتهم السياسية والاقتصادية والإعلامية. ومن أجل ذلك بدأت الحكومات باستخدام أنظمة تحليل النصوص من خلال أنظمة تحليل وسائل التواصل الاجتماعي (Social Media Analytics) التي سنتحدث عنها لاحقا.

٣. مستويات تحليل النصوص

هنالك ثلاثة مستويات أساسية لتطبيق تحليل النصوص تحدث عنها ريز (Ruiz 2009) وهي مستوى النص (Textual Level) والمستوى السياقي (Contextual

Level) والمستوى الاجتماعي (Sociological Level). تعمل المناهج المختلفة لأبحاث ومشاريع تحليل النصوص على مستوى أو أكثر من هذه المستويات.

في المستوى النصي، تُحَلَّل النصوص من حيث موضوعاتها (Topics) ومن حيث الجوانب المتعلقة ببنية وتركيب الخطاب (Discourse Composition and Structure). كذلك من خلال المستوى النصي يمكن اكتشاف أنماط (Patterns) محددة في النص نفسه يمكن الاستفادة منها تحليليا. أما التحليل على المستوى السياقي، فإنه يمكن أن يؤدي إلى اكتشاف معلومات ذات علاقة بسياق الخطاب (Discourse Context) أو السياق الاجتماعي (Register) الذي أُنتِج النص أو استُقبِل فيه (كالسياق الاجتماعي الذي كتبت فيه مشاركة على وسائل التواصل الاجتماعي). من جهة أخرى، في المستوى الاجتماعي من مستويات تحليل النصوص يُربَط النص الذي نقوم بتحليله بالمجال الاجتماعي الذي أنتج و استُقبِل فيه بعد تحليله على المستويين النصي والسياقي. في هذا المستوى يمكن أن يُحَلَّل النص كانعكاس لأيديولوجية الكاتب والمستقبِل كما يمكن أن يُحَلَّل النص كوسيلة لاستخلاص معلومات اجتماعية عن الكاتب أو المستقبِل.

٤. مراحل تحليل النصوص

٤, ١ اختيار حالة الدراسة

يتطلب تحليل النصوص اختيار حالة للدراسة (Case Selection) تتمثل في مجموعة من البيانات والوثائق المراد تحليلها واستخلاص النتائج والتعميمات منها. في هذا السياق، من أجل أن ينتج البحث نتائج أكثر شمولا حول ظاهرة معينة، يجب اختيار حالة ممثلة (Representative) أي حالة تمثل نسبة كبيرة من مجموعة اجتماعية معينة أو أن يتم اختيار عينة عشوائية تمثل تلك المجموعة الاجتماعية. إلى جانب ذلك، يذكر الباحثون في مجال تحليل النصوص أن هنالك حالات تعرف بالحالات الخاصة. هذه الحالات هي حالات لها أهمية استراتيجية إلا أنها ليست حالات عامة يمكن أن تمثل نسبة كبيرة من مجموعة اجتماعية معينة. من أمثلة الحالات الخاصة التي استُخدِمَ تحليل النصوص لدراستها البحث الذي نشره كل من جيبسون وزيلنيربرون (Gibson and Zellner-Bruhn, 2001) والذي قاما فيه بتحليل استخدام الموظفين في أربع دول للمجاز اللغوي (Metaphor). اختيرت الدول الأربع لأهميتها الاستراتيجية بحيث يمكن بعد استخلاص نتائج البحث حول

هذه الحالة الخاصة أن تُعَمَّم نتائج البحث على مساحات جغرافية أكبر لمعرفة الاختلاف الثقافي واللغوي في الدول الواقعة في تلك المساحات الجغرافية.

٤ , ٢ تحديد سؤال البحث أو المشروع

بعد تحديد حالة الدراسة، يبدأ الباحث في مجال تحليل النصوص بتحديد سؤال البحث أو المشروع (على سبيل المثال: هل ترتبط ظاهرة التبديل اللغوي (Language Shift) في حوارات وسائل التواصل الاجتماعي بجنس معين دون الآخر أو بتقسيمات ديموغرافية أخرى)؟

٤ , ٣ اختيار وجمع الوثائق والعينات النصية

بعد ذلك، يتم تحديد استراتيجية اختيار البيانات (Data Selection Strategy) والتي يتم من خلالها اختيار الوثائق أو مصادر البيانات التي ستُجمَع عينة النصوص (Text Sample) منها من أجل الإجابة عن سؤال البحث أو المشروع. من مصادر هذه البيانات نصوص مواقع معينة على الشبكة العنكبوتية. من هذه المصادر أيضا مشاركات وحوارات وسائل التواصل الاجتماعي كبيانات حوارات تويتر الحية أو التاريخية التي يوفرها أرشيف تويتر الرسمي (Twitter Gnip Firehose). كما يمكن استخدام مشاركات فيسبوك التاريخية التي يوفرها أرشيف Datasift. بعد تحديد استراتيجية اختيار البيانات، تأتي مرحلة تحديد استراتيجية جمع عينة البيانات (Data Sampling Strategy).

لعل القارئ يتساءل هنا، لماذا نحتاج لعينة البيانات؟ في كثير من الأحيان لا يستطيع الباحث جمع وتحليل كل البيانات النصية لمصدر معين. على سبيل المثال، يصعب على الباحث تحليل كل ما كتبته صحيفة معينة منذ إنشائها إلى الآن أو كل ما كُتِب من مشاركات حول موضوع معين على موقع من مواقع التواصل الاجتماعي منذ نشأته إلى الآن. في هذه الحالة يمكن استخدام استراتيجية لجمع عينة من البيانات. تجدر الإشارة هنا إلى أنه مع تطور تقنيات ومصادر البيانات الضخمة صار بالإمكان توفير مثل هذه العينات من البيانات بشكل أكثر دقة وسهولة. على سبيل المثال، عن طريق أرشيف عينات تويتر (Twitter Gnip Decahose) يمكن جمع عينة عشوائية تتكون من ١٠٪ من تغريدات تويتر الحية حول موضوع معين.

يجب أخذ العينة بشكل ممثِّل والمقصود بذلك هو أن تكون العينة ممثِّلة إحصائيا للمجموعة البشرية العامة التي أُخِذَت منها. تسمح العينة الممثلة لاحقا بتعميم نتائج البحث أو المشروع على مجموعة بشرية أكبر. في هذا السياق تعتبر العينة الاحتمالية (Probability Sample) العينة المثالية حيث يمكن عن طريق هذا النوع من العينات تعميم الاستنتاجات البحثية على مجموعة بشرية أكبر من خلال الاستدلالات الإحصائية. تجدر الإشارة هنا إلى أنه في مجال أبحاث ومشاريع تحليل النصوص هنالك صعوبات تواجه الباحث عند محاولته الحصول على عينة احتمالية ممثلة عند تحليل النصوص بشكل عام وعند تحليل النصوص المأخوذة من وسائل التواصل الاجتماعي بشكل خاص. تحدث كريبيندورف (Krippendorf 2013) عن هذه الصعوبات إلا أن المقام لا يتسع هنا للحديث عنها. لذلك لجأ الباحثون في مجال تحليل النصوص إلى استخدام استراتيجيات أخرى لجمع عينات النصوص منها استراتيجية التعداد (Enumeration) لوحدات النصوص كتعداد مشاركات وحوارات مواقع التواصل الاجتماعي على مدى سبعة أيام متتالية ثم الجمع العشوائي للعينات (Random Sampling) من هذه المشاركات والحوارات.

فمع تطور وزيادة حجم صفحات الشبكة العنكبوتية (The Web) ومع تطور أدوات معالجة البيانات الضخمة، صارت صفحات الشبكات العنكبوتية مصدرا من أهم مصادر البيانات النصية في أبحاث ومشاريع تحليل النصوص. في هذا السياق، هنالك أسلوبان لجمع مجموعات الوثائق والبيانات النصية من الشبكة العنكبوتية وهما أسلوب الزحف (Web Crawling) وأسلوب الكشط (Web Scrapping). يُطَبَّق أسلوب الزحف من خلال تحديد صفحات الشبكة العنكبوتية التي يجب أن تضاف إلى مجموعة البيانات المراد تحليلها عن طريق الإبحار في روابط الشابكة المرتبطة بهذه الصفحة (Link Navigation). يتم ذلك من خلال البدء بمجموعة أساسية من عناوين الشابكة (URLs) والتنقل عبرها للوصول إلى الصفحات المرتبطة بها. أما أسلوب الكشط فهو يتمثل في العملية التي تُستَخدَم لانتزاع (Extraction) النصوص

من مجموعة من صفحات الشبكة العنكبوتية والتي جُمِعَت عن طريق عملية الزحف (على سبيل المثال الصفحات المرتبطة بعنوان شابكة معين أو الصفحات التي تشكل موقعا معينا على الشبكة العنكبوتية).

إلى جانب صفحات الشبكة العنكبوتية التقليدية، تحتوي الشبكة العنكبوتية أيضا على مصادر بيانات نصية أخرى أفرزها ما يعرف بالجيل الثاني من الشبكة العنكبوتية (Web 2.0) والذي يشتمل على مواقع تتكون من محتوى يساهم بنشره المستخدمون مثل موقع ويكيبيديا و مواقع وسائل التواصل الاجتماعي مثل تويتر وفيسبوك ومدونات الشبكة العنكبوتية (Web Blogs). إضافة إلى ذلك، هنالك ما يعرف بالشبكة العنكبوتية العميقة (Deep Web) والتي تشتمل على بيانات مخزنة في قواعد للبيانات لاتصل إليها محركات البحث التقليدية.

تُستَخدَم أنظمة حاسوبية خاصة لجمع المعلومات النصية من الشبكة العنكبوتية باستخدام أسلوبي الزحف والكشط. يمكن تطبيق أسلوب الزحف باستخدام برمجيات جاهزة مفتوحة المصدر (Open Source) منها Nutch و Scraby. يمكن كذلك استخدام أحد برمجيات (أوامر) نظام التشغيل لينيكس (Linux) مثل wget والذي يسمح بتطبيق أسلوب الزحف بشكل آلي على أي مجموعة من عناوين الشابكة. من جهة أخرى يمكن استخدام برمجيات جاهزة تستطيع التعرف على أنواع متعددة من المحتويات في مواقع الشبكة العنكبوتية وانتزاع وتخزين أنواع البيانات التي يحددها المستخدم. يمكن أيضا استخدام لغات برمجة مثل Python لكتابة برامج تقوم بكشط البيانات (انظر ميتشيل ٢٠١٥، Mitchel). من أمثلة البرمجيات الجاهزة المستخدمة في كشط البيانات Helium Scraper و Outwit و FMiner. يمكن كذلك استخدام ما يعرف بواجهة برمجة التطبيقات (API) والتي تسمح بكشط البيانات من مواقع الشبكة العنكبوتية ومن وسائل التواصل الاجتماعي. يحتاج استخدام واجهات برمجة التطبيقات إلى توافر معرفة برمجية أساسية ولا يتطلب نفس المستوى من الخبرة البرمجية التي يتطلبها استخدام لغات البرمجة مثل Python لكتابة برامج كشط البيانات. من أمثلة واجهات برمجة التطبيقات المستخدمة لكشط النصوص من وسائل التواصل الاجتماعي واجهة برمجة تطبيقات تويتر (Twitter API) وأرشيف تويتر الرسمي (Twitter Gnip Firehose).

٤ , ٤ الصيغة المنطقية الاستدلالية

بعد الانتهاء من تحليل البيانات النصية التي جُمِعَت، يقوم الباحث باستخدام صيغة معينة من الصيغ المنطقية للحصول على استدلالات حول العلاقات التي تربط الظواهر التي دُرِسَت أو للحصول على استدلالات حول العلاقات التي تربط الظواهر المدروسة بالتعميمات النظرية. في مجال تحليل النصوص يمكن استخدام ما يعرف بالمنطق الاستقرائي (Inductive Logic) أو المنطق الاستنتاجي (Deductive Logic) للاستدلال والوصول إلى نتائج البحث أو المشروع.

٥. مصادر البيانات المعجمية الإلكترونية

تمثل مصادر البيانات المعجمية الإلكترونية (Lexical Resources) وسيلة إلكترونية لحفظ واسترجاع مجموعات كبيرة من البيانات المعجمية كالكلمات والمركبات (Phrases) مترافقة مع معلومات لغوية أخرى كالمعاني والعلاقات الدلالية. من أمثلة تلك المصادر الإلكترونية المعاجم الإلكترونية التي تشتمل على كلمات ومركبات مترافقة مع معاني تلك الكلمات والمركبات وطريقة استخدامها لغويا. من أمثلتها أيضا المكنز الإلكتروني (Thesaurus) والذي يصنف الكلمات المترابطة دلاليا والمترادفات في مجموعة واحدة. كذلك نجد من أمثلتها تلك المصادر التي تقوم بربط الكلمات والعبارات بحقول دلالية أو بدلالاتها الافتراضية من حيث المزاج العام (Default Sentiment) أو العاطفة (سلبي، إيجابي، محايد). كذلك من أمثلة تلك المصادر ما يعرف بقوائم الكلمات (Word Lists) وهي سجلات تشتمل على الصيغ الصرفية (Morphological Word Forms) الصحيحة الممكنة لكلمات لغة أو لهجة معينة.

في هذا السياق تمثل مصادر البيانات المعجمية مكونا هاما في أغلب تطبيقات تحليل النصوص كالتطبيقات التي تهتم بانتزاع المعلومات (Information Extraction) من خلال انتزاع أهم الكلمات والعبارات المستخدمة في نصوص معينة، والتطبيقات التي تهتم باكتشاف العلاقات بين الكلمات في النصوص. كذلك تُستخدَم في تطبيقات تحليل النصوص المتعلقة بتصنيف النصوص (Text Classification) وتحليل المزاج العام (Sentiment Analysis). من جهة أخرى يعتمد تصميم أنظمة التدقيق الإملائي (Spell Checkers) للنصوص على قوائم الكلمات (Word Lists) لتكوين المعجم الإلكتروني لتلك الأنظمة من أجل معرفة التهجئة الصحيحة للكلمات.

تجدر الإشارة إلى أن تصميم المصادر المعجمية يتطلب وقتا وجهدا كبيرا يمتد إلى سنوات. كما يتطلب خبراء لغويين متخصصين في صناعة المعاجم (Lexicography). كذلك يتطلب الأمر في بعض المصادر الاستعانة بمتخصصين في علم النفس والمنطق. من أبرز أمثلة المصادر المعجمية المستخدمة عالميا قاعدة البيانات المعجمية WordNet وهي شبكة إلكترونية دلالية بدأ بتصميمها جورج ميلر في العام ١٩٨٥ في جامعة برينستون الأمريكية. تحوي هذه القاعدة أغلب الأسماء والأفعال والصفات والأحوال في اللغة الإنجليزية.

يمثل هذا النظام قاعدة بيانات معجمية تحوي الكلمات ومعانيها ومرادفاتها وعلاقاتها المعجمية. هذه القاعدة مبنية على نظريات علم اللغة النفسي (Psycholinguistics) المتعلقة بتمثيل المعرفة المعجمية في الذاكرة المعجمية (Lexical Memory) الإنسانية. تحوي هذه القاعدة ما يعرف بمجموعات الترادف (Synsets). كل مجموعة ترادف تحوي مجموعة من الكلمات المترادفة التي تمثل مفهوما معجميا (Lexical Concept) أساسيا. تحتوي قاعدة بيانات WordNet على ١٥٥ ألف كلمة مصنفة إلى ١١٧ ألف مجموعة ترادف وتُربَط مجموعات الترادف ببعضها عن طريق علاقات دلالية.

٦. المعالجة الحاسوبية للنصوص

بعد تحديد الوثائق والعينات النصية وجمع تلك الوثائق والعينات النصية من مصادر النصوص، تأتي مرحلة المعالجة الحاسوبية للنصوص (Text Processing). قبل التمكن من تحليل نص أو مجموعة من النصوص، يجب معالجة هذه النصوص لجعلها قابلة للتحليل واستخلاص النتائج. من أمثلة هذه المعالجات، إزالة وسوم (Tags) لغة HTML المستخدمة في وسم وترميز صفحات الشبكة العنكبوتية وذلك في حال استخدام نصوص تم الحصول عليها من مصادر تعتمد على الشبكة العنكبوتية وإزالة وسوم لغة XML المستخدمة في تخزين الوثائق الإلكترونية. من عمليات المعالجة أيضا، تقسيم النص إلى الكلمات (Tokenization)، إزالة علامات الترقيم الملتصقة بالكلمات، حذف كلمات الإيقاف (Stop Words)، تجريد الكلمات إلى جذوعها (/ Stemming Lemmatization)، إحصاء الكلمات ، الوسم النحوي للكلمات (Part of Speech Tagging)، وربط الكلمات بمعانيها المعجمية.

٦ , ١ تقسيم النص إلى كلمات (Tokenization)

في هذه العملية يتعرف الحاسوب على الكلمات في النص باعتبار أن المسافات وعلامات الترقيم هي حدود فاصلة بين الكلمات. كما يقوم أيضا في هذه العملية بحذف أدوات الترقيم الملتصقة بالكلمات. على سبيل المثال، في تتابع الكلمات التالي "الطقس في هذا اليوم جميل!"، يقوم الحاسوب بتقسيم هذا التتابع إلى الكلمات التالية: الطقس، في، هذا، اليوم، جميل. نلاحظ هنا أن الحاسوب قام بتقطيع النص إلى كلمات منفصلة مع حذف علامة التعجب الملتصقة بالكلمة الأخيرة. تُنتِج هذه العملية من النص مجموعة من الكلمات (Tokens) يمكن استخدامها في تحليل هذا النص أو تطبيق عمليات إحصائية عليه. كما يمكن أن تُستخدم هذه المجموعة من الكلمات كمدخلات (Inputs) لتطبيقات أخرى مثل تطبيقات التحليل الصرفي (Morphological Analysis) أو تحليل المزاج العام أو تصنيف النصوص. تجب الإشارة هنا إلى أنه في كثير من التطبيقات المرتبطة بمعالجة النصوص (مثل تطبيقات تحليل وسائل التواصل الاجتماعي)، نحتاج إلى تنقية (Filtering) مجموعة الكلمات المستخرجة من نص معين عن طريق حذف الكلمات الوظيفية (Function Words) كحروف الجر والضمائر المنفصلة والظروف وغيرها من الكلمات الوظيفية التي ترد بشكل كبير في النصوص وتعرف في مجال تحليل النصوص بكلمات الإيقاف (Stop Words) . لذلك، تُستَخدَم قائمة من كلمات الإيقاف التي يرجع إليها الحاسوب من أجل تنقية مجموعة الكلمات التي استخلصها من نص معين. يقوم الحاسوب بتنقية مجموعة الكلمات المستخلصة عن طريق حذف الكلمات الموجودة في قائمة كلمات الإيقاف لتبقى في المجموعة الكلمات ذات المحتوى (Content Words) وهي الكلمات ذات الأهمية في تحليل النصوص كالأسماء والأفعال.

٦ , ٢ استخلاص جذع الكلمة (Stemming/Lemmatization)

عملية استخلاص جذع الكلمة (Stemming) هي عملية يُستَخلَصُ فيها الجزء الأساسي من الكلمة المشتقة (Derived) أو المصرفة (Inflected) بعد حذف السوابق (Prefixes) واللواحق (Suffixes) من الكلمة. على سبيل المثال الكلمات: كاتِبان، كاتِبات، كاتِبا، الكاتِبان، للكاتِبَيْن لها جذع (Stem) أساسي واحد وهو كاتِب. يساعد

استخلاص جذع الكلمة في تحديد العلاقات بين الكلمات المترابطة صرفيا أو دلاليا مع اختلافها في البنية السطحية (Surface Structure) . ينبغي أن نشير هنا إلى أن عملية استخلاص جذع الكلمة لا يقصد بها إرجاع الكلمة إلى الجذر أو تجريد الكلمات من حروف الزيادة كما هو متعارف عليه في الدراسات الصرفية التقليدية.

تجب الإشارة هنا أيضا إلى أن عملية استخلاص جذع الكلمة من خلال إزالة السوابق واللواحق قد لا تكون كافية لإرجاع بعض الكلمات إلى جذعها الأساسي، حيث إن إزالة السوابق واللواحق من كلمة مشتقة أو مصرفة قد تُنتِج جذعا غير مستخدم لغويا (أي ليس موجودا في معجم اللغة) أو قد يرجع أكثر من كلمة مرتبطة دلاليا وصرفيا إلى أكثر من جذع مع أنها في الأصل تشترك في جذع أساسي واحد. على سبيل المثال، في اللغة الإنجليزية، عندما تُطَبَّق عملية استخلاص جذع الكلمة على كلمة مثل having فإنها ستنتج لنا جذعا غير مستخدم لغويا وهو hav وذلك من خلال إزالة اللاحقة {ing-}. كذلك في اللغة العربية، قد نجد أنفسنا أمام كلمات جُمِعت جمع تكسير وهي كلمات لا نستطيع فقط الاعتماد على إزالة السوابق واللواحق منها لاستخلاص جذعها. على سبيل المثال، الكلمات: الطالب، طالبان، الطالبات، الطلاب كلها تعود إلى جذع واحد وهو طالب. نلاحظ هنا أننا في كلمة الطلاب (جمع تكسير) نستطيع أن نرجعها إلى جذعها المفرد المستخدم لغويا وهو طالب وهو نفس جذع الكلمات الأخرى المرتبطة بها دلاليا وصرفيا (الطالب، طالبان، الطالبات) دون الاعتماد فقط على إزالة السواق واللواحق. حيث إننا لو اعتمدنا فقط على إزالة السوابق واللواحق، فسنحصل على جذعين مختلفين لهذه الكلمات المترابطة دلاليا وصرفيا؛ حيث سنحصل على الجذع طالب للكلمات الطالب، طالبان، الطالبات والجذع طلاب لكلمة الطلاب (بعد حذف السابقة {ال-}).

لمواجهة ذلك، تُستخدَم عملية أخرى وهي عملية استخلاص الصيغة الصرفية الأساسية للكلمة (Lemmatization). الصيغة الصرفية الأساسية للكلمة (Lemma) هي أصغر صيغة للكلمة مُستخدَمة لغويا أي صيغة الكلمة دون وجود لسوابق أو لواحق تصريفية أو اشتقاقية أو ضمائر متصلة بشرط أن تكون هذه الصيغة الصرفية مُستخدَمة لغويا (موجودة في معجم اللغة). تقابل هذه الصيغة في اللغة العربية صيغة الماضي المفرد المذكر الغائب للأفعال وصيغة المفرد المذكر النكرة للأسماء. على سبيل

المثال: كَتَب، استكتَب، كاتِب، مكتوب، كِتاب، مَكتب تمثل صيغا صرفية أساسية (Lemmas) لكلمات مشتقة من جذر واحد (ك ت ب).

٦, ٣ إحصاءات النصوص

بعد تقسيم النص إلى كلمات واستخلاص جذع الكلمة أو استخلاص الصيغة الصرفية الأساسية للكلمة، نحصل على مجموعة من الكلمات يمكن أن نطبق عليها عمليات إحصائية تبين لنا أكثر الكلمات استخداماً في نص معين. كذلك يمكن أن يُستخدَم التحليل الإحصائي لمعرفة متتابعات الكلمات الأكثر استخداماً (n-grams) مثل أكثر كلمتين متتابعتين (Bi-Grams) أو أكثر ثلاث كلمات متتابعة (Tri-Grams) أو أكثر أربع كلمات متتابعة (Quad-Grams) استخداما. يمكن الاستفادة من إحصاءات أكثر متتابعات الكلمات استخداماً في تحديد التراكيب الأكثر استخداماً في نص معين. في هذا السياق تُعد إحصاءات الكلمات وإحصاءات متتابعات الكلمات من أهم عمليات تحليل النصوص المستخدمة في أنظمة تحليل وسائل التواصل الاجتماعي والأنظمة المستخدمة في التحليل الحاسوبي للمدونات اللغوية (Corpus Processing). حيث تُستخدم هذه العملية للتعرف على أكثر الموضوعات التي تحدث عنها نص معين إضافة إلى تحديد الكلمات المفتاحية (Keywords) والعبارات المفتاحية (Key Phrases) المستخدمة في النص.

٦, ٤ وسم الفئة النحوية للكلمات (Part of Speech Tagging)

هي العملية التي يقوم خلالها الحاسوب بمسح نص معين ووسم (Tagging) كل كلمة فيه بالفئة النحوية (Syntactic Category) التي تنتمي إليها (اسم، فعل، حال، صفة ...) بناء على معنى هذه الكلمة في المعجم وبناء على السياق الذي تأتي فيه هذه الكلمة (علاقتها مع الكلمات أو العبارات الأخرى التي تأتي معها في السياق). على سبيل المثال، يستطيع الحاسوب وسم كلمات هذه الجملة "أكل الولد التفاحة" كالتالي:

أكل	ال	الولد	ال	تفاحة
Verb	Determiner	Noun	Determiner	Noun

تجدر الإشارة هنا إلى أن هنالك اختلافاً في الوسوم المستخدمة في أنظمة وسم الفئة النحوية للكلمات حيث تقوم بعض الأنظمة باستخدام وسوم بسيطة مثل Verb (فعل) و Noun (اسم)، بينما تقوم أنظمة أخرى باستخدام وسوم أكثر تركيبا وتفصيلا كما هو الحال في الوسوم المستخدمة في مدونة Penn Treebank ، وهي عبارة عن مدونة نصية خضعت للتحليل النحوي وحُدِّدَت الفئات والعلاقات والأبنية النحوية لجملها. حيث استُخدمت في هذه المدونة وسوم مركبة مثل:NN والتي تدل على اسم عام مفرد، و NNSوتدل على اسم عام جمع، و NNP وتدل على اسم علم مفرد، وNNPS وتدل على اسم علم جمع.

تعتمد أنظمة وسم الفئة النحوية للكلمات غالبا على ما يعرف بالتعلم الإشرافي (Supervised Learning) والذي يعتمد على المعرفة التي يستخلصها الحاسوب من نصوص حُدِّدَت فئاتها النحوية سابقا لتعلم كيفية وسم الفئة النحوية لكلمات نصوص أخرى آليا.

تُستخدَم عملية وسم الفئات النحوية للكلمات في معالجة النصوص لتحديد التراكيب النحوية الأكثر استخداماً في نص معين بشكل دقيق من خلال التعرف على أنماط استخدام التركيب الإضافي وغيره من التراكيب النحوية. كذلك يُستخدم وسم الفئات النحوية في التحليل النحوي الآلي للنصوص (Syntactic Parsing) والذي يُمَكِّن من تحديد العلاقات النحوية بين الوحدات النحوية (Syntactic Constituents) للجمل في النصوص وتمثيل البنية النحوية لتلك الجمل كتحديد المركبات الاسمية (Noun Phrases) التي تتكون من أداة تحديد يتلوها اسم أو المركبات الفعلية (Verb Phrases) التي تتكون من فعل يتلوه مركب اسمي. لزيادة دقة التحليل النحوي الآلي، تَستخدِم أنظمة التحليل النحوي التعلم الإشرافي من خلال تدريب النظام على نصوص تم تحليلها نحويا بشكل يدوي كنصوص مدونة Penn Treebank لتعلم كيفية تحليل جمل نصوص أخرى بشكل آلي بناء على المعرفة المستخلصة من النصوص المُحلَّلة يدويا.

٦,٥ وسم أسماء الكيانات (Named Entity Tagging)

في هذه العملية يتعرف الحاسوب على أسماء الكيانات (Named Entities) مثل أسماء الأشخاص والمواقع والبلدان والشركات سواء كانت كلمات مفردة أو تعبيرات مركبة من أكثر من كلمة. تدخل هذه العملية في إطار ما يعرف بعملية انتزاع المعلومات (Information Extraction).

٦,٦ النماذج اللغوية

النماذج اللغوية (Language Models) هي نماذج احتمالية (Probabilistic) تَستخدم نظرية الاحتمالية (Probability Theory) للتعامل مع اللغة الطبيعية بحيث يمكن استخدامها لتَوَقُّع احتمالية استخدام تتابع معين من الكلمات أو الحروف في نصوص لغة معينة. كذلك تستخدم هذه النماذج في تحديد مدى احتمالية (Likelihood) ترافق (Collocation) كلمتين في نصوص لغة معينة (على سبيل المثال ترافق كلمتي "التنمية" و "الاقتصادية").

لبناء النماذج اللغوية تُستَخدَم المدونات اللغوية التي تتميز بحجم نصوصها الكبير. تُستخدَم هذه المدونات اللغوية كمدونات تدريبية (Training Corpora) لتدريب الحاسوب من أجل بناء النموذج اللغوي، وذلك من خلال حساب احتماليات تتابع كلمات أو حروف معينة في مجموعة من النصوص. تزداد دقة النموذج اللغوي كلما زاد حجم مجموعات النصوص التي يُبنى عليها النموذج. على سبيل المثال، عندما تُحتَسَب احتمالية الترافق لكلمتي "التنمية" و "الاقتصادية" بناء على نص مأخوذ من كتاب واحد فإن دقة احتمالية الترافق ستكون أقل بكثير من دقة الاحتمالية المُستَنتَجة من مدونة مكونة من نصوص صحف إلكترونية على مدى خمس سنوات.

تُستخدَم النماذج اللغوية في عدد من تطبيقات المعالجة الحاسوبية للغة الطبيعية، ومنها أنظمة التدقيق الإملائي والتعرف الآلي على الكلام المنطوق (Speech Recognition) والترجمة الآلية (Machine Translation) . تُستخدَم النماذج اللغوية أيضا في تحليل النصوص لتحديد التراكيب الأكثر استخداماً في نص معين أو لزيادة دقة تحديد المزاج العام لنص معين من خلال توقع ترافق كلمات معينة مع كلمات أخرى في النص بحيث يؤدي هذا الترافق إلى إعطاء الجملة مزاجا إيجابيا مثلا أو يؤدي ذلك الترافق إلى عكس المزاج الافتراضي للجملة من سلبي إلى إيجابي.

٦,٧ برمجيات المعالجة الحاسوبية للنصوص

لتطبيق المعالجة الحاسوبية للنصوص يمكن استخدام برمجيات جاهزة ومكتبات برمجية مفتوحة المصدر. من أبرز هذه البرمجيات أداة CoreNLP Toolkit وهي حزمة برمجية مكتوبة بلغة جافا طُوِّرَت في جامعة ستانفورد الأمريكية وتوفر أدوات

لتحليل النصوص (http://stanfordnlp.github.io/CoreNLP). كذلك يمكن استخدام أداة(GATE General Architecture for Text Engineering) التي طُوِّرَت في جامعة شيفيلد البريطانية وهي أيضاً حزمة برمجية مكتوبة بلغة جافا وتوفر أدوات لمعالجة النصوص (انظر ريز ٢٠١٥، Reese). كذلك يمكن استخدام أداة (LingPipe http://alias-i.com/lingpipe) وهي حزمة برمجية مكتوبة بلغة جافا توفر أدوات للمعالجة الحاسوبية للغة الطبيعية بشكل عام ومعالجة النصوص بشكل خاص (انظر بالدوين وآخرون ٢٠١٤، Baldwin). يمكن أيضا استخدام مكتبة البرمجيات(NLTK Natural Language Toolkit) وهي عبارة عن مكتبة برمجيات مصممة للاستخدام عن طريق لغة Python للبرمجة لأغراض معالجة النصوص (http://www.nltk.org) (انظر بيرد وآخرون ٢٠٠٩، Bird et. al.).

٧. تطبيقات تحليل النصوص

٧,١ تصنيف النصوص (Text Classification)

تطبيقات تصنيف النصوص هي تطبيقات يقوم الحاسوب من خلالها بإعطاء نص معين تصنيفاً أو أكثر من مجموعة من التصنيفات المحددة مسبقا. في هذا السياق يمكن عن طريق هذه العملية تصنيف الوثائق وفقا للموضوع، اللغة، الكاتب، أو غير ذلك من التصنيفات.

في البداية، كانت أنظمة تصنيف النصوص تعتمد على استخدام قوانين مركبة يستخدمها نظام التصنيف لتحديد تصنيف وثيقة أو نص معين بناء على تواجد كلمات معينة في هذا النص. على سبيل المثال، يمكن استخدام القانون التالي: إذا وُجِدَت كلمة "معلم" و "مَدْرَسة" و "اختبار" في نص واحد فإن هذا يعني أنه يمكن تصنيف هذا النص في المجال التعليمي أو التربوي. إلا أن أنظمة التصنيف المبنية على القوانين (Rule-Based) واجهتها صعوبة بناء وصيانة قوانين التصنيف وهي صعوبات تزداد بازدياد حجم النصوص المطلوب تصنيفها وبالتالي زيادة القوانين المطلوبة للتصنيف وهو ما يعرف بالقابلية للتوسع (Scalability). بناء على ذلك طُوِّرَت أنظمة لتصنيف النصوص تستخدم تقنية تعلم الآلة (Machine Learning). تعتمد هذه الأنظمة على التعلم الإشرافي، بحيث تُستَخدَم نصوص مصنفة سابقاً كبيانات لتدريب نظام التصنيف

عليها، وبذلك يتعلم الحاسوب –على سبيل المثال– أن كلمتي Free و Gift غالبا ما تترافقان في محتوى نصوص ما يعرف برسائل البريد الإلكتروني الإقحامية (Spam) بعكس كلمتي Research و Abstract اللتين تترافقان غالبا في محتوى نصوص رسائل غير إقحامية. تحتاج هذه الأنظمة إلى توافر نصوص مصنفة مسبقاً للتعلم منها، وبالتالي متى ما استطاع نظام التصنيف تعلم التصنيف من تلك النصوص المصنفة مسبقا فإنه سيستطيع تطبيق ما تعلمه لتصنيف أي نصوص أخرى، وبالتالي يكون النظام قابلا للتوسع (Scalable) بشكل أفضل.

من أمثلة تطبيقات تصنيف النصوص أنظمة التعرف على رسائل البريد الإلكتروني الإقحامية (Spam E-Mail Detection). في هذه الأنظمة يقوم الحاسوب بشكل آلي بتصنيف كل رسالة بريد إلكتروني مُستَلَمَة إلى مجموعة الرسائل الإقحامية (Spam) أو مجموعة الرسائل غير الإقحامية. تعمل هذه الأنظمة على مستوى الحاسوب الخادم لنظام البريد الإلكتروني (E-Mail Server) أو على مستوى التطبيق الذي يُستخدم لإرسال واستقبال وإدارة رسائل البريد الإلكتروني (E-Mail Client Application). حيث يقوم التطبيق بفحص رسائل البريد الإلكتروني الواردة إلى صندوق بريد المستخدم ليقرر تصنيفها إلى رسالة إقحامية لتصل إلى مجلد الرسائل الإقحامية (Spam/Junk Folder) أو تصنيفها إلى رسالة غير إقحامية لتصل إلى مجلد صندوق البريد الوارد (Inbox).

كذلك من أمثلة تطبيقات تصنيف النصوص التطبيقات الخاصة بتصنيف الموضوعات (Topic Classification) وهي عملية تُصَنَّف فيها الوثائق إلى موضوعات مثل الاقتصاد، السياسة، والطب. تُستخدَم هذه العملية عادة لتصنيف صفحات الشبكة العنكبوتية. حيث استُخدِمَت هذه العملية في مشروع الدليل المفتوح (Open Directory Project) لصفحات الشبكة العنكبوتية (http://www.dmoz.org) والذي صُنِّفَت من خلاله ملايين من صفحات الشبكة العنكبوتية إلى فئات شجرية بناء على موضوع الصفحة.

من جهة أخرى، تُستخدَم تطبيقات تصنيف النصوص في نوع آخر من الأنظمة التي تقوم بالتعرف على أسلوب الكاتب (Author Profiling). في هذا النوع من الأنظمة يقوم الحاسوب بتحديد معلومات وسِمات عن كاتب نص معين مثل العمر والجنس والاتجاه السياسي (كوبيل و آخرون ٢٠٠٢، Koppel et. al). في هذا السياق، يستخدم

الحاسوب –على سبيل المثال– معلومات متعلقة بإحصاءات استخدام أنواع معينة من الكلمات في النص كالكلمات الوظيفية (Function Words) للتعرف على أسلوب الكاتب أو جنسه.

٢,٧ انتزاع المعلومات (Information Extraction)

في هذه العملية يقوم الحاسوب باستخلاص معلومات ذات بنية منظمة (Structured Data) من مجموعة بيانات لا تسير وفق بنية منظمة (Unstructured Data) (إيجناتو وميخاليكا ٢٠١٦، Ignatow & Mihalcea). في هذه الحالة تُستَخلَص أنواع محددة مسبقا من البيانات مثل أسماء الأشخاص أو البلدان أو الشركات أو المنتجات من مجموعة من النصوص. كذلك يمكن استخلاص أحداث (Events) معينة كارتفاع أسعار أسهم شركة معينة أو عملة معينة. كما يمكن استخلاص العلاقات (Relations) مثل علاقة "رئيس شركة". لتوضيح الفكرة، يمكن لأنظمة انتزاع المعلومات أن تعالج النص التالي "سيقوم تيم كوك الرئيس التنفيذي لشركة أبل بإطلاق الإصدار الجديد لجهاز آي فون في مؤتمر أبل الذي سيعقد في سان فرانسيسكو في السابع من الشهر الجاري". من هذا النص يمكن لنظام انتزاع المعلومات أن يستخلص أسماء كيانات (Entity Names) وتتمثل في اسم شخص (Person Name) وهو تيم كوك، واسم منظمة/ شركة (Organization Name) وهو أبل، وحدث (Event) وهو مؤتمر أبل، ووقت وهو السابع من الشهر الجاري. كذلك يمكن للنظام أن يستخلص من هذا النص علاقة وهي علاقة "الرئيس التنفيذي" بين تيم كوك و شركة أبل. هنا نجد أن معلومات اسم الشخص، اسم المنظمة/ الشركة، الحدث، الوقت، والعلاقة التي استخلصها الحاسوب هي بيانات ذات بنية منظمة استُخلِصَت من نص يمثل بيانات لا تسير وفق بنية منظمة.

يمكن الاستفادة من عملية انتزاع المعلومات في تطبيقات كثيرة؛ حيث توفر هذه العملية إمكانية الحصول على بيانات وتنظيمها لتحليلها ومعالجتها لاحقا (Post Processing). حيث يمكن على سبيل المثال استخدام عملية انتزاع المعلومات للحصول على بيانات حول منتجات معينة أو شركات معينة أو أسعار أسهم معينة أو علاقات معينة كرؤساء شركات أو أساتذة جامعات. يمكن أن تُدخَل هذه البيانات إلى حقول قاعدة بيانات للاستفادة منها لاحقا ولربطها مع بيانات أخرى.

في هذا السياق، هنالك عمليتان أساسيتان يقوم عليهما عمل أنظمة انتزاع المعلومات وهما انتزاع الكيانات (Entity Extraction) انتزاع العلاقات (Relations Extraction).

٧, ٢, ١ انتزاع الكيانات

من أجل تحقيق الاستفادة من أنظمة انتزاع المعلومات يجب أن يستطيع الحاسوب التعرف على أسماء الكيانات (Named Entities) المذكورة في نص معين. يدخل في هذا المجال التعرف على أسماء الأشخاص، أسماء البلدان، أسماء الشركات، أسماء المنتجات، وأسماء الأحداث. كذلك يجب أن يستطيع الحاسوب التعرف على الكيانات التي تُصَنَّف تحت مجموعات دلالية معينة مثل الحيوانات والأطعمة.

لتحقيق انتزاع الكيانات، يُستَخدَم عادة معجم يحوي أسماء كيانات مثل أسماء شركات وأسماء أشخاص بحيث يرجع الحاسوب إلى هذا المعجم للتعرف على أسماء الكيانات الموجودة في نص معين. في هذا السياق، يمكن على سبيل المثال أن يحوي هذا المعجم أسماء مثل Boeing و Samsung و Sony و Apple و IBM كأسماء لشركات. كذلك يمكن استخدام مجموعة من النصوص التي وُسِمَت يدويا بأسماء كيانات (Tagged Named Entities) ليستخلص الحاسوب منها معجما يحوي أسماء الكيانات في هذه المجموعة من النصوص. باستخدام هذا المعجم الذي بُنِيَ من تلك النصوص الموسومة، يمكن للحاسوب التعرف آليا على أسماء الكيانات في نصوص أخرى غير موسومة (Unannotated Texts). كذلك يستطيع الحاسوب - بناءً على هذا المعجم - استنتاج وتَعلُّم قوانين تحدد له إذا ما كان أمام اسم كيان. على سبيل المثال، يستطيع الحاسوب استنتاج أن كلمتي "شركة" أو "مؤسسة" تأتيان في سياق الاستخدام اللغوي قبل اسم كيان لشركة أو مؤسسة. كذلك يمكن أن يستنتج أنه إذا وجد نمطا (Pattern) مثل "أنا موظف في" فإنه غالبا ما يكون الاسم الذي يأتي بعد النمط اسم كيان يدل على شركة أو مؤسسة أو منظمة يعمل فيها موظفون. على سبيل المثال، عندما يجد الحاسوب جملة مثل "أنا موظف في رويترز" فإنه سيضيف اسم "رويترز" إلى معجمه كاسم كيان يدل على منظمة/ شركة. من خلال هذه القوانين والأنماط، يستطيع الحاسوب التعرف على المزيد من أسماء الكيانات وإضافتها إلى معجم أسماء الكيانات لديه. وبهذا يزداد حجم المعجم ويُستخدَم مرة أخرى بالترافق مع القوانين للتعرف على أسماء كيانات جديدة في

نصوص جديدة؛ وكلما تعرف الحاسوب على أسماء كيانات جديدة، يقوم بإضافة أسماء الكيانات الجديدة تلك إلى معجم أسماء الكيانات بشكل تكراري (Recursively). تُعرَف هذه العملية في مجال انتزاع المعلومات بعملية (Bootstrapping) وهي عملية تقوم على استخدام قائمة بأسماء الكيانات وقائمة بالأنماط أو القوانين لِتَعَلُّم المزيد من أسماء الكيانات من نصوص بشكل تصاعدي. استُخدِمَت هذه الطريقة أيضا للتعرف على أسماء كيانات لفئات دلالية (ريلوف وجون ١٩٩٩، Riloff & Jone).

٧, ٢, ٢ انتزاع العلاقات

كثيرا ما نحتاج إلى معرفة العلاقة التي تربط بين شخصين أو شخص ومنظمة معينة أو غير ذلك من العلاقات. على سبيل المثال، عندما نقول إن الشخص س هو أخ الشخص ص فإن هنالك علاقة تربط بين الشخص س والشخص ص وهي علاقة أخوة. وعندما نقول إن الشخص س يعمل في الشركة ص فإن هنالك علاقة تربط بين الشخص س والشركة ص وهي علاقة عمل. تُعرَف العملية التي تُمَكِّن الحاسوب من معرفة العلاقات بين الكيانات بعملية انتزاع العلاقات (Relations Extraction) وهي أحد مجالات انتزاع المعلومات.

قبل التعرف على العلاقات التي تربط كيانات معينة يجب أولا التعرف على تلك الكيانات ثم بعد ذلك الانتقال إلى تحديد العلاقات التي تربطها. إلا أن عملية التعرف على العلاقات التي تربط الكيانات تعتبر أكثر صعوبة من عملية التعرف على الكيانات.

كما رأينا في عملية التعرف على الكيانات، لتمكين الحاسوب من التعرف على العلاقات تُستَخدَم عادة مجموعة من النصوص الموسومة يدويا بعلاقات تربط الكيانات الموجودة فيها. بعد ذلك تُستَخدَم تقنية تعلم الآلة لتمكين الحاسوب من التدرب والتعلم على تحديد هذه العلاقات في نصوص أخرى. يتطلب تحديد العلاقات أيضا استخدام سمات تميز الكيانات التي تشترك في علاقة معينة كسمة الجنس المستخدمة في علاقة أخ أو أخت. كذلك تُستخدَم سمات أخرى متعلقة بالدور الدلالي (Thematic Role) للكيانات أو الدور النحوي لها (من خلال استخدام التحليل النحوي (Parsing) وأشجار التحليل (Parse Trees). باستخدام كل هذه السمات والبيانات، يستطيع الحاسوب تحديد العلاقات التي تربط بين الكيانات في النصوص.

٣,٧ استرجاع المعلومات (Information Retrieval)

في هذه العملية يقوم الحاسوب بمعالجة استفسار (Query) لطلب بيانات موجودة في وثائق أو صفحات على الشبكة العنكبوتية أو على أنظمة إدارة الوثائق (Document Management Systems) ثم التعرف على الوثائق أو الصفحات التي تحوي المعلومات المطلوبة في الاستفسار واسترجاعها. يعتمد الحاسوب في تحديد الوثائق أو الصفحات المطلوبة على مستوى التشابه بين الاستفسار ومحتوي النص الموجود في الوثائق أو الصفحات التي يبحث فيها لتحديد ما إذا كانت هذه الوثائق أو الصفحات هي المطلوبة في الاستفسار.

ليتمكن الحاسوب من إنجاز ذلك، تُعالَج أولا نصوص الوثائق والصفحات معالجة أولية (Pre-Processing). تشتمل هذه المعالجة الأولية على تطبيق عملية تقسيم النص إلى كلمات (Tokenization) واستبعاد كلمات الإيقاف (Stop Words) واستخلاص جذع الكلمة (Stemming/Lemmatization) (ارجع إلى القسم السادس من هذا المبحث). بعد ذلك يقوم الحاسوب بفهرسة (Indexing) البيانات النصية الموجودة في الوثائق أو الصفحات المراد البحث فيها. خلال عملية الفهرسة يُكَوِّن الحاسوب فهرسا وهو بنية بيانات (Data Structure) تقوم بإجراء عملية ربط (Mapping) بين الكلمات من جهة والوثائق أو الصفحات المُفهرَسة من جهة أخرى بحيث يحدد الفهرس الوثائق أو الصفحات التي توجد فيها كل كلمة في ذلك الفهرس. يجب أن تُبنى بنية الفهرس بشكل يُمَكِّن الحاسوب من الوصول بشكل سريع إلى الوثائق أو الصفحات التي تحوي كل كلمة في ذلك الفهرس. بعد ذلك يستقبل الحاسوب استفسارات من المستخدم يقوم على ضوئها بالبحث عن المعلومات المطلوبة بالرجوع إلى الفهرس واسترجاع الوثائق أو الصفحات المتعلقة بالاستفسارات. في هذا السياق، عندما يقوم المستخدم بإدخال استفسار إلى نظام استرجاع المعلومات، فإن النظام يرجع إلى الفهرس للبحث عن الكلمات المفتاحية (Keywords) الموجودة في الاستفسار ويسترجع الصفحات أو الوثائق التي تحوي هذه الكلمات المفتاحية بناء على ما وجده في الفهرس. عندما يسترجع النظام الوثائق أو الصفحات ذات العلاقة بالاستفسار، فإنه يعرض للمستخدم نتائج البحث مرتبةً وفقاً لدرجة ارتباط هذه الوثائق بكلمات الاستفسار (Relevance) بحيث تكون الوثائق الأقرب لكلمات الاستفسار أعلى في الترتيب (Ranking) من الوثائق

الأقل ارتباطا. تُطَبَّق هذه العمليات من خلال واجهة استخدام (User Interface) تقوم باستقبال الاستفسار الذي يُدخله المستخدم ثم استرجاع الوثائق أو الصفحات ذات الصلة بالاستفسار، وترتيبها وفقا لدرجة الارتباط بذلك الاستفسار. كذلك يمكن أن تقوم هذه الواجهة بتحسين دقة الاستفسار الذي يُدخِله المستخدم عن طريق التدقيق الإملائي للاستفسار والتصحيح الآلي (Automatic Correction) للكلمات المُدخَلة من خلال عرض النتائج المقابلة للصيغة الصحيحة لغويا للكلمة التي أدخلها المستخدم بشكل خاطئ إلى النظام. على سبيل المثال، عندما يقوم المستخدم بإدخال الاستفسار الخاطئ التالي "إظراب العمال" فإن واجهة الاستخدام لنظام استرجاع المعلومات تقوم آليا بتصحيح الاستفسار وتحويله إلى الصيغة الصحيحة إملائيا "إضراب العمال" واسترجاع الوثائق والصفحات المرتبطة بهذه الصيغة الصحيحة. كذلك يُمكِن من خلال واجهة الاستخدام عرض نتائج التحليل المرادفة للاستفسار (Synonyms) باستخدام مكنز مرادفات آلي (Thesaurus). كذلك يمكن من خلال هذه الواجهة إكمال الاستفسار المُدخَل بالكلمات التي عادة ما تترافق مع الجزء الذي أدخله المستخدم مثل إكمال الاستفسار المكون من كلمة "الصحة" بكلمة "النفسية" ليصبح الاستفسار "الصحة النفسية".

استُخدِمَت أنظمة استرجاع المعلومات في عدد من المجالات أهمها محركات البحث (Search Engines) على الشبكة العنكبوتية وتطبيقات البحث في أنظمة إدارة الوثائق. حيث تُستخدَم أنظمة استرجاع المعلومات عادة للبحث في نصوص لا تسير وفق بنية منظمة، مثل صفحات الشبكة العنكبوتية والمحتوى النصي الكامل (Full Text) لوثائق مخزنة في أنظمة إدارة الوثائق. إلا أن هذه الأنظمة تستخدم أيضا للبحث في بيانات تخضع لبنية منظمة كالبيانات المخزنة في قواعد البيانات، كقواعد بيانات الصور والأفلام وسجلات أوراق الأبحاث العلمية.

لتحديد ما إذا كانت وثيقة أو صفحة معينة لها علاقة بالاستفسار الذي أدخله المستخدم، يقوم الحاسوب بالبحث عن مكونات الاستفسار لمحاولة العثور على ما يقابله في الوثائق أو الصفحات التي يبحث فيها. نشير هنا إلى أنه في كثير من الحالات لا يمكن أن تتطابق كل مكونات الاستفسار مع الوثائق أو الصفحات التي يبحث الحاسوب فيها. لذلك تلجأ أنظمة استرجاع المعلومات إلى البحث عن الكلمات

الموجودة في الاستفسار بشكل حر دون التزام بالترتيب الذي كُتِبَت به هذه الكلمات في الاستفسار وهو ما يعرف بأسلوب حقيبة الكلمات (Bag-of-Words). على سبيل المثال، عندما يكون الاستفسار مكوناً من التتابع التالي من الكلمات "اللسانيات الحاسوبية اللغة العربية" فإن الحاسوب يستطيع أن يسترجع أي وثائق أو صفحات تحمل هذه الكلمات دون الالتزام بترتيبها الذي كُتِبَت به في الاستفسار. في هذه الحالة سيسترجع الحاسوب وثائق أو صفحات تحوي التتابع التالي "اللغة العربية في ضوء دراسات اللسانيات الحاسوبية" أو "أبحاث اللسانيات الحاسوبية في مجال معالجة اللغة العربية". في هذا السياق، تَستخدِم أنظمة استرجاع المعلومات نماذج معيارية للبحث عن معلومات الاستفسار أهمها نموذج بوليان (Boolean Model) و ونموذج فيكتور سبيس (Vector Space) والنموذج الاحتمالي (Probabilistic Model). سنتحدث هنا باختصار عن هذه النماذج حيث إن المقام في هذا المبحث لا يتسع للحديث عنها بالتفصيل.

نموذج بوليان هو نموذج مبني على المنطق الرمزي (Symbolic Logic) حيث يَبحث الحاسوب باستخدام هذا النموذج عن الكلمات المفتاحية المستخدمة في الاستفسار الذي يُدخِله المستخدم من خلال دمج الكلمات المفتاحية مع أدوات البحث المنطقية (Logical Search Operators) مثل AND (و)، OR (أو)، و NOT (ليس) للوصول إلى نتائج البحث. على سبيل المثال، عندما يقوم المستخدم بإدخال الاستفسار التالي "العربية AND اللسانيات" فإن الحاسوب سيبحث عن الوثائق أو الصفحات التي تحوي كلتا الكلمتين المفتاحيتين "العربية" و "اللسانيات"؛ أما عندما يُدخل المستخدم الاستفسار التالي "العربية OR اللسانيات" فإن الحاسوب سيبحث عن الوثائق أو الصفحات التي تحوي إحدى الكلمتين "العربية" أو "اللسانيات". إلى جانب نموذج بوليان، يُستخدَم نموذج فيكتور سبيس وهو نموج يعتمد على حساب مدى التشابه بين حقيبة الكلمات (Bag-of-Words) المُكَوِّنة لاستفسار المستخدم ونصوص الوثائق أو الصفحات التي يبحث فيها الحاسوب. كما يُستخدَم أيضا النموذج الاحتمالي الذي يعتمد على تحديد مدى احتمالية مطابقة وثيقة أو صفحة معينة للاستفسار الذي أدخله المستخدم باستخدام تطبيقات نظرية الاحتمالية (Probability Theory) بدلا من الاعتماد فقط على الكلمات المفتاحية المُدخَلَة في الاستفسار.

٧, ٤ تحليل المزاج العام (Sentiment Analysis)

من تطبيقات تحليل النصوص أيضا أنظمة تحليل المزاج العام (Sentiment Analysis) وتُعرَف أيضا بأنظمة تحليل الرأي (Opinion Mining). لقيت هذه الأنظمة اهتماما واسعا في السنوات الأخيرة مع تطور وزيادة استخدام أنظمة تحليل وسائل التواصل الاجتماعي (Social Media Analytics) والتي تشتمل على تحليل المزاج العام كأحد مكوناتها (انظر وانج وآخرون ٢٠١٢، Wang, H. et. al). في هذه الأنظمة يقوم الحاسوب بتصنيف نص معين من حيث مزاج المحتوى إلى إيجابي أو سلبي أو محايد. في هذا السياق، توفر أنظمة تحليل المزاج العام للجهات الحكومية وسيلة لقياس تفاعل الجمهور وردود أفعالهم تجاه الأحداث والقضايا السياسية والاجتماعية والاقتصادية وغيرها. كذلك يُمكِن استخدام تحليل البيانات التاريخية لوسائل التواصل الاجتماعي لمعرفة التغير التاريخي في اتجاه المزاج العام تجاه قضية معينة (Trend Analysis and Benchmarking). كذلك تَستخدِم الشركات أنظمة تحليل المزاج العام لمعرفة آراء الزبائن في منتجاتهم و خدماتهم من خلال تحليل مزاج مشاركات وحوارات وسائل التواصل الاجتماعي التي تتناول منتجاتهم.

إلى جانب استخدامه في تحليل وسائل التواصل الاجتماعي، استُخدِمَ تحليل المزاج العام في تطبيقات أخرى لتحليل لنصوص حيث استُخدِمَ في تحليل المزاج العام للأخبار (انظر لويد وآخرون ٢٠٠٥، Lloyd, L et. al.) و تحليل رأي مستخدمي مواقع تقييم المنتجات (Product Reviews) (انظر هو وآخرون ٢٠٠٤، Hu, M et. al.).

٧, ٤, ١ مصادر البيانات المعجمية لتحليل المزاج العام

من أجل تحديد المزاج العام للنصوص، نحتاج إلى المرور بأكثر من مرحلة من مراحل تحليل النصوص. في المرحلة الأولى يُبنَى معجم إلكتروني (Lexicon) يحوي مجموعة كبيرة من الكلمات والعبارات التي حُدِّدَ المزاج العام الافتراضي (Default Sentiment Polarity) المقابل لها بشكل يدوي (على سبيل المثال، كلمة "رائع" تحمل المزاج "إيجابي" بينما كلمة "سيء" تحمل المزاج "سلبي"). يَستخدِم الحاسوب هذا المعجم بعد ذلك لتحديد المزاج العام لنصوص وسائل التواصل الاجتماعي وغيرها من مصادر النصوص. تجدر الإشارة إلى أنه في بعض السياقات قد لا تكون قيمة المزاج

العام الافتراضية لكلمة واحدة في الجملة كافية لتحديد المزاج العام لكل الجملة. على سبيل المثال، الجملة التالية "أنا لا أحب التفاح" تُعتبر سلبية من حيث المزاج على الرغم من كون كلمة "أحب" إيجابية من حيث المزاج العام. إلا أنها في هذا السياق (بعد لا النافية) تعطي قيمة عكسية للمزاج العام، فيصبح المزاج العام للجملة سلبيا. لذلك في بعض الأحيان تحتاج أنظمة تحليل المزاج العام إلى تحليل سياق الجملة لمعرفة المزاج العام لها دون الاعتماد فقط على المزاج العام لكلمة من كلماتها. من جهة أخرى، قد تكون الجملة إيجابية من حيث المزاج العام، إلا أن كاتبها يقصد في استخدامها التهكم أو السخرية أو قد يقصد توجيه إسقاطات أو انتقادات ذات طبيعة سياسية مثلا بشكل غير مباشر. في هذه الحالة تصبح الجملة سلبية من حيث المزاج العام على الرغم من أن المزاج العام الظاهر لهذه الجملة إيجابي. بناء على ذلك، ومن أجل زيادة دقة نتائج التحليل الآلي للمزاج العام، يُمزَج التحليل الحاسوبي الآلي للمزاج العام مع المراجعة البشرية. حيث يُحلَّل المزاج العام أولاً بشكل آلي باستخدام نظام تحليل المزاج العام. بعد ذلك تُستَخدَم المراجعة البشرية لنتائج التحليل لتنقيحها. تجدر الإشارة إلى أن هذه المراجعة البشرية تساعد في الحصول على نتائج أكثر دقة لتحليل المزاج العام، حيث إن المراجعة البشرية لا تنظر فقط إلى المعنى الظاهر للكلمات والعبارات بل تستخدم المعرفة البشرية المتعلقة بجوانب الدلالة والسياق والأبعاد ذات العلاقة بالسياسة، الاقتصاد، التقاليد وغيرها من الجوانب التي تؤثر في الحكم على المزاج العام لجملة معينة.

من أمثلة المعاجم الإلكترونية المستخدمة في تحليل المزاج العام المعجم المرافق لنظام OpinionFinder (انظر ويبي وآخرون ٢٠٠٥، Wiebe, et. al.). حيث طورت مجموعة من الباحثين من جامعات بيتسبيرج و كورنيل ويوتاه الأمريكية نظام (OpinionFinder http://mpqa.cs.pitt.edu/opinionfinde) والذي يقوم بمعالجة الوثائق النصية والتعرف الآلي على جوانب الرأي الشخصي في جمل هذه الوثائق وتحديد المزاج العام لها. بُنِيَ المعجم الإلكتروني المرافق لنظام OpinionFinder باستخدام كلمات جُمِعَت وحُدِّدَ مزاجها العام يدويا، إضافة إلى استخدام كلمات استُخرِجَت من مدونات لغوية وُسِمَ المزاح العام لكلماتها مسبقا. يحتوي المعجم على ٦٨٥٦ مادة معجمية منها ٩٩٠ مادة مكونة من تعبيرات مركبة (Multiword Expressions). إضافة إلى تحديد المزاج العام (إيجابي، سلبي، محايد) لكل مادة معجمية في المعجم، حُدِّدَت الفئة النحوية

(Part of Speech) لهذه المواد المعجمية. تجدر الإشارة إلى أن نظام OpinionFinder استُخدِم لوسم جزء من جمل مدونة Penn Treebank وهي كما أشرنا سابقا عبارة عن مدونة نصية حُلِّلَت نحويا وحُدِّدَت الفئات والعلاقات والأبنية النحوية لجملها.

كذلك من أمثلة المعاجم الإلكترونية المستخدمة في تحليل المزاج العام معجم SentiWordNet والمستخدم في تحليل الرأي (Opinion Mining) (انظر إيسولي وآخرون ٢٠٠٦،.Esuli, A. et. al). بُنِيَ هذا المعجم اعتمادا على قاعدة البيانات المعجمية WordNet، حيث يعطي هذا المعجم لكل مجموعة ترادف (Synset) في قاعدة بيانات WordNet قيمة تبين المزاج العام لهذه المجموعة (إيجابي، سلبي، محايد). لتكوين معجم SentiWordNet تم البدء بوسم عدد من مجموعات الترادف يدويا بقيمة المزاج العام المقابل لها. بعد ذلك وُسِمَت مجموعات الترادف الأخرى بشكل آلي ليغطي معجم SentiWordNet الآن كل مجموعات الترادف في قاعدة WordNet المعجمية، حيث وصل عدد كلمات المعجم إلى ١٠٠ ألف كلمة.

من جهة أخرى استخدم الباحثون مدونات لغوية مكونة مجموعات من نصوص وُسِمَت كلماتها بالمزاج العام المقابل لها. من خلال تَعَلُّم الحاسوب من الكلمات الموسومة وسياقاتها في نصوص تلك المدونات اللغوية، يقوم الحاسوب آليا بتحديد المزاج العام لنصوص آخري أو تكوين معاجم إلكترونية لتُستَخدَم في تحليل المزاح العام (يعرف ذلك باستخدام التعلم الإشرافي لتحديد المزاج العام).

من أمثلة المدونات اللغوية الموسومة مسبقا بِقِيَم المزاج العام لكلماتها مدونة Multiperspective Question Answering وهي عبارة عن مدونة لغوية تحوي ٥٣٥ مقالة إخبارية باللغة الإنجليزية جُمِعَت من مصادر إخبارية متنوعة ووُسِمَت يدويا بمعلومات حول المزاج العام والعواطف التي تعبر عنها عباراتها (انظر ويبي وآخرون ٢٠٠٥، Wiebe, et. al).

من جهة أخرى، صُمِّمَت مدونات لغوية متخصصة في مجال آراء مشاهدي الأفلام (Movie Review) واستُخدِمَت في تدريب الحاسوب على تحليل المزاج العام في نصوص أخرى في مجال آراء المشاهدين في الأفلام وفي مجالات أخرى. في هذا السياق، قام كل من بانج و لي (انظر بانج وآخرون ٢٠٠٤،.Pang, B. et. al) بتصميم مدونة مكونة من مجموعتي نصوص: الأولى تحوي ألف مقالة لآراء المشاهدين وتحمل مزاجا

إيجابيا والثانية تحوي نفس العدد من المقالات وتحمل مزاجا سلبيا. كذلك قام ماس وآخرون (انظر ماس وآخرون ٢٠١١،.Maas, A. L. et. al) بتكوين مدونة أكبر متخصصة في مجال آراء مشاهدي الأفلام تحتوي على ٥٠ آلف مقالة لآراء المشاهدين جُمِعَت من موقع IMDb المتخصص في الأفلام.

تجدر الإشارة إلى أن الباحثين في مجال تحليل المزاج العام وتحليل الرأي بدأوا بالاهتمام بالمواقع التي تحوي مقالات ومشاركات تعبر عن آراء المستخدمين في منتجات معينة (Products Review) مثل موقع amazon.com و epinions.com حيث استُخدِمَ المحتوى النصي لهذه المواقع لتكوين مدونات لتحليل المزاج العام يمكن استخدامها لبناء معاجم إلكترونية لتحليل المزاج العام، فضلا عن استخدامها في مجال تعلم الحاسوب للتحليل السياقي للمزاج العام لنصوص أخرى.

٧ , ٤ , ٢ أنظمة تحليل المزاج العام

يمكن تقسيم أنظمة تحليل المزاج العام إلى فئتين، الفئة الأولى هي الأنظمة المبنية على القوانين (Rule-based) وهي أنظمة تعتمد في عملها على معاجم إلكترونية بُنِيَت بشكل يدوي أو آلي. النوع الثاني من هذه الأنظمة يتمثل في الأنظمة التي تعتمد على تقنية تَعلُّم الآلة (Machine Learning) وهي أنظمة تقوم بتحليل المزاج العام للنصوص من خلال المعرفة التي تَدَرَّبَ الحاسوب عليها عن طريق المدونات الموسومة بمعلومات المزاج العام. من أمثلة أنظمة النوع الأول التي تعتمد على القوانين نظام OpinionFinder السابق ذكره. حيث يقوم هذا النظام بالتعرف الآلي على المزاج العام لكلمات وعبارات النصوص التي يحللها وفقا لوجود أو عدم وجود تلك الكلمات والعبارات في معجمه الإلكتروني. من أمثلة النوع الثاني الذي يعتمد في تحليله للمزاج العام على تقنية تَعلُّم الآلة نظام تحليل المزاج العام للوثائق الذي طوره كل من بانج و لي (بانج وآخرون ٢٠٠٤،.Pang, B. et. al). لعل ما يميز الأنظمة التي تعتمد على التعلم الآلي عن طريق مدونات لغوية موسومة مسبقا بمعلومات المزاج العام أن هذه الأنظمة يمكن استخدامها لتحليل المزاج العام لأي لغة متى ما توافرت نصوص المدونات اللغوية الموسومة بمعلومات المزاج العام التي يتعلم منها الحاسوب. كذلك ظهرت فئة حديثة من أنظمة تحليل المزاج العام وهي الأنظمة المعتمدة على التعلم الآلي العميق (Deep Learning) والذي يتمثل في استخدام التعلُّم الآلي من خلال ما يُعرَف ببنوك

أشجار تحليل المزاج العام (Sentiment Treebanks)، حيث تَستخدِم هذه الأنظمة معلومات المزاج العام للكلمات والعبارات بالترافق مع أشجار التحليل النحوي (Parse Trees) لتحليل المزاج العام للجمل في النصوص بشكل تركيبي (انظر سوشر وآخرون ٢٠١٣،.Socher, R. et. al). يساعد ذلك في التعامل مع الجمل المركبة التي لا تسير وفقا لمزاج واحد بل تحوي تغيرا في المزاج مثل جملة "تصميم المنزل رائع ولكن نوافذه قبيحة". حيث نلاحظ أن هذه الجملة المركبة بدأت بمزاج إيجابي (تصميم المنزل رائع) إلا أن المزاج تحول في النصف الثاني من الجملة إلى مزاج سلبي (ولكن نوافذه قبيحة). لمعرفة المزيد حول استخدام التعلم الآلي العميق في مجال تحليل المزاج العام، يمكن للقارئ زيارة هذه الصفحة من موقع جامعة ستانفورد الأمريكية http://nlp.stanford.edu/sentiment.

٨. الخاتمة

من خلال اكتشاف وانتزاع معرفة هامة من نصوص حرة لا تسير وفق بنية منظمة، يظهر دور تحليل النصوص كتطبيق هام من تطبيقات المعالجة الحاسوبية للغة الطبيعية. يتحقق ذلك من خلال التفاعل بين أكثر من مجال بيني أهمها علم الحاسوب، اللسانيات الحاسوبية، استرجاع المعلومات، تحليل البيانات، تعلم الآلة، والإحصاء. يسير استخدام تحليل النصوص جنبا إلى جنب مع التطور الكبير في مجال البيانات الضخمة والذي تسبب بإنتاج كميات هائلة من البيانات النصية، وإيجاد تطبيقات ومنصات تحليلية عديدة، ولغات برمجة وأدوات برمجية وخوارزميات متخصصة للتعامل مع هذا الكم الهائل من البيانات النصية. من خلال تطبيقات تحليل النصوص يمكن الاستفادة من تلك البيانات للوصول إلى نتائج علمية ومؤشرات ذات أهمية وفائدة للباحثين ومتخذي القرار في الجهات الحكومية والتجارية. كل ذلك يبين أهمية اللسانيات الحاسوبية في عصر المعلومات كتخصص علمي له أبعاد علمية وتطبيقية في شتى مجالات الحياة.

المراجع

◇ **Baldwin, Breck, and Krishna Dayanidhi**. Natural language processing with Java and LingPipe Cookbook. Packt Publishing Ltd, 2014.

◇ **Bird, Steven, Ewan Klein, and Edward Loper**. Natural language processing with Python. " O'Reilly Media, Inc.", 2009.

◇ **Chakraborty, Goutam, Murali Pagolu, and Satish Garla**. Text mining and analysis: practical methods, examples, and case studies using SAS. SAS Institute, 2014.

◇ **Esuli, Andrea, and Fabrizio Sebastiani**. "Sentiwordnet: A publicly available lexical resource for opinion mining." Proceedings of LREC. Vol. 6. 2006.

◇ **Gibson, Cristina B., and Mary E. Zellmer-Bruhn**."Metaphors and meaning: An intercultural analysis of the concept of teamwork." Administrative Science Quarterly 46.2 (2001): 274-303.

◇ **Hardeniya, Nitin**. NLTK essentials. Packt Publishing Ltd, 2015.

◇ **Hu, Minqing, and Bing Liu**. "Mining and summarizing customer reviews." Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2004.

◇ **Ignatow, Gabe, and Rada Mihalcea**. Text Mining: A Guidebook for the Social Sciences. SAGE Publications, 2016.

◇ **Khan, Aamera ZH, Mohammad Atique, and V. M. Thakare**. "Combining lexicon-based and learning-based methods for Twitter sentiment analysis." International Journal of Electronics, Communication and Soft Computing Science & Engineering (IJECSCSE) (2015): 89.

◇ **Koppel, Moshe, Shlomo Argamon, and Anat Rachel Shimoni**. "Automatically categorizing written texts by author gender." Literary and Linguistic Computing 17.4 (2002): 401-412.

◇ **Krippendorff, Klaus**. Content analysis: An introduction to its methodology. Sage, 2012.

◇ **Lloyd, Levon, Dimitrios Kechagias, and Steven Skiena.** "Lydia: A system for large-scale news analysis." International Symposium on String Processing and Information Retrieval. Springer Berlin Heidelberg, 2005.

◇ **Maas, Andrew L., Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Pott**. "Learning word vectors for sentiment analysis." Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1. Association for Computational Linguistics, 2011.

◇ **Mihalcea, Rada**. "Using Wikipedia for Automatic Word Sense Disambiguation." HLT-NAACL. 2007.

◇ **Mitchell, Ryan**. Web scraping with Python: collecting data from the modern web. " O'Reilly Media, Inc.", 2015.

◇ **Pang, Bo, and Lillian Lee.** "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts." Proceedings of the 42nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004.

◇ **Perkins, Jacob**. Python 3 Text Processing with NLTK 3 Cookbook. Packt Publishing Ltd, 2014.

◇ **Reese, Richard M.** Natural language processing with Java. Packt Publishing Ltd, 2015.

◇ **Riloff, Ellen, and Rosie Jones**. "Learning dictionaries for information extraction by multi-level bootstrapping." AAAI/IAAI. 1999.

◇ **Ruiz, Jorge Ruiz**. "Sociological discourse analysis: Methods and logic."Forum Qualitative Sozialforschung/Forum: Qualitative Social Research. Vol. 10. No. 2. 2009. Retrieved August 26, 2016, from http://www.qualitative-research.net/index.php/fqs/article/view/1298/2882.

◇ **Shaw, Jonathan**. "Why "Big Data" is a big deal." Harvard Magazine 3 (2014): 30-35. Retrieved August 22, 2016, from http://harvardmagazine.com/2014/03/why-big-data-is-a-big-deal.

◇ **Socher, Richard, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts.** "Recursive deep models for semantic compositionality over a sentiment treebank." Proceedings of the conference on empirical methods in natural language processing (EMNLP). Vol. 1631. 2013.

◇ **Stamatatos, Efstathios, Nikos Fakotakis, and George Kokkinakis.** "Automatic text categorization in terms of genre and author." Computational linguistics 26.4 (2000): 471-495.

◇ **Strapparava, Carlo, and Rada Mihalcea.** "Learning to identify emotions in text." Proceedings of the 2008 ACM symposium on Applied computing. ACM, 2008.

◇ **Wang, Hao, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan**. "A system for real-time twitter sentiment analysis of 2012 us presidential election cycle." Proceedings of the ACL 2012 System Demonstrations. Association for Computational Linguistics, 2012.

◇ **Wiebe, Janyce, Theresa Wilson, and Claire Cardie**. "Annotating expressions of opinions and emotions in language." Language resources and evaluation 39.2-3 (2005): 165-210.

الفصل السادس

التدقيق الإملائي

د. وليد بن عبدالله الصانع[1]

ملخص البحث

يعتبر الحاسب الآلي الآن هو الأداة الرئيسية التي يستخدمها المؤلفون في الكتابة. ونظرا لأهمية الكتابة السليمة لإيصال الفكرة، فإن المدقق الإملائي يعتبر عنصرا مهما في برمجيات التحرير الكتابي على أجهزة الحاسب الآلي. وقد دأبت كبرى شركات التقنية منذ ظهور الحاسب الآلي على تطوير مدققات إملائية تساعد المؤلفين على اكتشاف وتصحيح الأخطاء الإملائية. وتعتبر اللغة العربية الآن من اللغات المدعومة في كثير من أنظمة التشغيل وأجهزة الحاسب الآلي والبرمجيات. وقد قامت كبريات الشركات العالمية بتطوير مدققات إملائية للغة العربية. ونظرا لأن اللغة العربية مستخدمة في بقاع كثيرة، وهي تعتبر من اللغات القديمة والتي ما زال كثير من عباراتها المستخدمة قديما دارجة الاستخدام في المؤلفات الحديثة، فإن تطوير مدققات إملائية لها يعتبر تحديا لا

١-يعمل الدكتور وليد بن عبدالله الصانع أستاذ بحث مساعد بالمركز الوطني لتقنية الحاسب والرياضيات التطبيقية بمدينة الملك عبدالعزيز للعلوم والتقنية. حصل على درجتي البكالوريوس والماجستير في علوم الحاسب الآلي من جامعة الملك سعود. قرأ الدكتوراه في مجموعة الذكاء الاصطناعي بجامعة يورك ببريطانيا. عمل سابقا مهندسا للبرمجيات في قسم الأبحاث والتطوير في شركة الإلكترونيات المتقدمة ومهندسا للنظم والبرمجيات في شركة الاتصالات السعودية. تتمثل اهتماماته البحثية في تعليم الآلة وتحديدا في البرامج المنطقية، الرسوم الاحتمالية، البرامج المنطقية الاحتمالية وتطبيقات هذه النظريات في نمذجة ومعالجة اللغة والأدوات التعليمية.(walsanie@kacst.edu.sa)

بأس به بسبب اختلاف صيغ الإملاء عبر الزمن وبين البقاع. وفي هذا الفصل أقوم أولا باستعراض أبرز التحديات التي تواجه مطوري المدققات الإملائية للغة العربية. ومن ثم أعرج على طرق اكتشاف الأخطاء الإملائية وإشكالياتها. ومن ثم أستعرض أبرز الطرق لتصحيح الأخطاء الإملائية. وفي نهاية الفصل، أعطي نبذة سريعة عن بعض النظريات المتقدمة التي تستخدم في أبحاث تطوير المدققات الإملائية مؤخرا وبعض المراجع الأساسية التي من الممكن أن يرجع إليها القارئ.

الفصل السادس: التدقيق الإملائي

١. تمهيد

منذ بداية نظرية الحوسبة في منتصف القرن الميلادي الماضي ومعالجة اللغات الطبيعة كانت وما تزال واحدة من أهم المجالات التي يهتم بها الباحثون في مجالات الحوسبة واللغويات والرياضيات والفلسفة. وتكمن أهمية معالجة اللغات الطبيعية في أن اللغة هي الوسيلة التي يستخدمها الإنسان للتعبير عن مشاعره وللتواصل مع الآخرين، لذا كان من الأهمية أن يتم العمل على حوسبة اللغة ليتمكن الإنسان من التواصل مع الحاسوب دون الحاجة لتعلم لغة جديدة. فبرز مجال الذكاء الاصطناعي كأحد أهم المجالات التي تم العمل عليها منذ ظهور نظرية الحوسبة. وكانت معالجة اللغات الطبيعية أبرز ملامح هذا الفن. فكان تأسيس مجال الذكاء الاصطناعي، والمتمثل بالاختبار الافتراضي الذي وضعه ألان تيورنج (Alan Turing) والمسمى بـ اختبار تيورنج (Turing Test)، معتمدا على التواصل مع الآلة باستخدام اللغة البشرية المكتوبة (Turing، ١٩٥٠، Russell و Norving، ١٩٩٥، الفصل ٢٦). حيث إن مبدأ الاختبار يقوم على أن يتم وضع آلة وإنسان في مكان غير مرئي. ثم يقوم شخص بالحوار مع الآلة والإنسان من خلال الكتابة، بحيث يكتب السؤال ويرسله إلى أي من الاثنين ومن ثم تأتيه الإجابة

مكتوبة. فإذا كان هذا المحاور لا يستطيع التفريق بين الآلة والإنسان من خلال الردود التي تأتيه فإن الآلة تعتبر حينئذ ذكية وتكون قد اجتازت الاختبار. فكما هو ملاحظ من خلال هذه المنهجية أن معالجة اللغة تعتبر ركيزة في تأسيس مجال الذكاء الاصطناعي. ولا غرابة في ذلك، فاللغة تعتبر من أهم الوسائل للحصول على المعرفة وإيصالها[١].

و لما كانت النصوص المكتوبة تمثل الطريقة الأمثل والأسهل لتخزين المعرفة في الحاسوب، برزت أهمية العمل على تطوير البرمجيات التي تساعد المؤلف على الكتابة السليمة الخالية من الأخطاء. لذا بدأ العمل على تطوير المدقق الإملائي. فمنذ ستينيات القرن الميلادي الماضي، عكف الباحثون على فهم طرق الكتابة والتأليف وتطوير المعاجم اللغوية التي تساعد على اكتشاف الأخطاء الإملائية وتصحيحها (Damerau, 1964). ومن ثم قامت كبريات شركات التقنية، كأي بي إم (IBM) وإي تي أند تي (AT&T)، بتطوير هذه البرمجيات. وقد درج كثير من الباحثين على تطوير نظريات اكتشاف الأخطاء وتصحيحها ومن ثم تطويعها للعمل على بيئات أكثر تعقيدا، كالنصوص المكتوبة من غير المتقنين، والأخطاء الإملائية التي تنتج كلمات أخرى صحيحة ولكنها ليست الكلمات المبتغاة.

وفي هذا الفصل سأقوم بداية بإعطاء نبذة عن مشكلة الإملاء في اللغة العربية وبعض التحديات التي يواجهها الباحثون في بناء المدقق الإملائي. وسأقوم أيضا في الجزء الأول من هذا الفصل بعرض لمحة عن المبادئ الرئيسة التي يعتمدها كثير من الباحثين في بناء المدقق الإملائي. ومن ثم في الجزء الثاني من هذا الفصل أستعرض طرق اكتشاف الأخطاء الإملائية والتحديات التي تعتريها. وفي الجزء الثالث أستعرض الطرق الأساسية التي طرحها الباحثون لتصحيح الأخطاء الإملائية آليا. ولأن الهدف من هذا الفصل هو إعطاء نبذة عن المدققات الإملائية فإنني سأتجنب طرق التصحيح المتقدمة لكي يكون الفصل متاحا للقراء من مختلف الخلفيات المعرفية. وفي نهاية هذا

١-مصادر المعرفة ثلاثة:

- الاستنتاج الذاتي Prior knowledge
- الإحساس Perception
- النقل Testimony

و اللغة هي الوسيلة التي تستخدم في الحصول على المعرفة من المصدر الأخير (النقل).

الفصل أذكر بشكل سريع ومختصر الفكرة الأساسية لطرق التدقيق المتقدمة المبنية على نظرية الاحتمالات ومن ثم أسرد بعض المراجع التي يمكن للقارئ أن يرجع إليها إذا أراد الاستزادة من هذا الباب.

٢. التدقيق الإملائي للغة العربية

٢, ١ اللغة العربية وإشكاليات قواعد الإملاء

كما هو معلوم لدى الكثيرين أن طرق الكتابة في اللغة العربية مرت بمراحل متعددة. فقد كانت الكتابة في بداية عصور التدوين تتم بحروف غير منقوطة. ولم تكن الهمزة وحروف التشكيل معروفة لدى العرب. فكانت الأحرف تعرف من سياق الكلمة. فمثلا حرفي ال "ح" وال "ج" لهما نفس الرسم "ح". لذا فكلمتي "رحل" و"رجل" تكتبان "رحل" ويكون التمييز بينها أثناء القراءة من خلال السياق. وأثناء عصور التدوين ظهرت الحاجة للتفريق بين الأحرف التي لها نفس الرسم، فبدأ استخدام التنقيط كوسيلة لذلك. وتم أيضا فيما بعد إدخال حركات التشكيل والهمزة للتمييز بين الكلمات التي لها نفس الأحرف الهجائية ولكن تختلف من حيث النطق. ونظرا لتوسع رقعة العالم الإسلامي، فإن هذه الإضافات أحدثت بعض الاختلافات في قواعد كتابتها. وتغيرت هذه القواعد مع مرور الزمن وحدثت اختلافات في طريقة الكتابة بين أقاليم العالم الإسلامي. فالياء المتطرفة (في نهاية الكلمة) مثلا، تكتب في كتب التراث الإسلامي بلا تنقيط. أما في العصر الحديث فإنها تكتب بشكل شائع في كثير من دول العالم العربي بالنقط ولكن ظلت تكتب في بعض الدول العربية، كمصر مثلا، بلا تنقيط جريا على ما كانت عليه كتب التراث. كذلك بالنسبة للهمزة المتوسطة، فقد ظهرت طرق مختلفة لكتابتها بين بعض الأقاليم. فكلمة "مسؤولية" مثلا، تكتب بالطريقة السابقة وتكتب أيضا بهذه الطريقة "مسئولية". فهذه الاختلافات في طرق الكتابة تجعل اعتبار ما هو صواب وما هو خطأ من حيث الإملاء أمرا غير محسوم.

وبالرغم من أن اللغة العربية التي كتب بها القرآن الكريم محفوظة بحفظ الله لها في كتابه، إلا أن اللغة العربية الدارجة على الألسن تتأثر بما يعتريها مع تغير الأزمنة، فتدخلها كلمات لم تكن معروفة أو شائعة على الألسن في زمن ما. لذا فإن بعض الدارسين للغة

العربية يصنفون اللغة العربية إلى صنفين[١]:

- الصنف الأول ويتمثل في العربية الفصيحة التقليدية (Classical Arabic)، وهي اللغة المستخدمة في كتب التراث.
- الصنف الثاني ويتمثل في العربية الفصيحة الحديثة (Modern Standard Arabic)، وهي اللغة المستخدمة في الكتابات الرسمية في العصر الحديث.

إذ إن العربية الفصيحة الحديثة، فيما يرى المؤمنون بهذا التصنيف، تختلف عن العربية الفصيحة التقليدية من حيث الألفاظ، نظرا لدخول كلمات جديدة كما ذكرنا سابقا أو قلة استخدام بعض الكلمات التي كانت شائعة في وقت ما. وتختلف أيضا من حيث الإملاء، وقد ذكرت مثالا سابقا يتمثل في حالتي الهمزة والتنقيط. هذا فضلا عن الاختلافات في الظواهر اللغوية الأخرى التي لا تهمنا في هذا الفصل. وهذه التغيرات تجعل من التدقيق الإملائي عملا أكثر تحديا. فمعالجة النصوص التراثية تختلف عن معالجة النصوص الحديثة. ومعالجة نصوص مكتوبة في أحد الأقاليم قد تختلف عن معالجة نصوص مكتوبة في إقليم آخر.

و لعل من أبرز صعوبات التدقيق الإملائي أيضا تلك المتعلقة بالتعرف على أسماء الأعلام. ولمناقشة الصعوبات المتعلقة بأسماء الأعلام، أستعرض حالتين. الحالة الأولى هي تغير نطق الأسماء مع الزمن، وهذا يؤثر على طريقة كتابتها. أما الحالة الثانية فهي دخول أسماء جديدة لم تكن معروفة في وقت سابق. ولكي نستعرض مثالا على الحالة الأولى نأخذ الاسم "سارة". فإنه يكتب حاليا في نصوص متعددة بهذه الطريقة "سارا" نظرا لنطق الاسم بلا تحريك في الغالب، فتنطق التاء المربوطة في آخره هاء، والتي تتلقاها بعض المسامع ألفا؛ لتشابه مخرجي الحرفين الهاء والألف. إضافة إلى أن الشكل الأعجمي للاسم "Sara" يكون فيه الصوت الأخير حرفا يوافق الألف لا التاء أو الهاء، فيستخدم بعض الكتاب في العصر الحديث الصيغة المقابلة للنطق الأعجمي للاسم. أما الحالة الثانية، وهي دخول أسماء جديدة على اللغة، فإنه عند المرور بكلمة غريبة مثلا، ومن دون معرفة السياق، فإنه لا يمكن التأكد مما إذا كانت هذه الكلمة اسم

١- الجدير بالذكر أن كثيرا من متخصصي اللغة العربية يرفضون هذا التصنيف. ولست هنا بصدد ترجيح رأي أي من الفريقين، ولكن أعرض هذا التصنيف كحالة موجودة في الدراسات اللغوية الحديثة، وتحديدا في دراسات الحوسبة اللغوية.

علم أو كلمة أخرى كتبت بطريقة خاطئة. وخذ هذه الجملة مثالا توضيحيا:

أعطيتها لانا.

و هي جملة صائبة حيث إنها تعني أن المتحدث أعطى فتاة اسمها "لانا" شيئا ما، فلو كان الاسم "لانا" غير موجود في المعجم الذهني للقارئ فإنه من دون الرجوع إلى الجمل السابقة أو اللاحقة لهذه الجملة، والتي يمكن من خلالها معرفة أن الكاتب يتحدث عن شخص ما هنا وبالتالي يدخل هذا الاسم الجديد إلى معجمه الذهني، فإن القارئ إذا أخذ هذه الجملة بمعزل عن السياق فإنه ربما اعتبر هذه الكلمة خطأ إملائيا، ورجح أن الكاتب أراد أن يقول «أعطيتها لبنا» ولكنه قلب الباء ألفا.

لذا فإن هذه الأمور التي ذكرتها تمثل بعض التحديات التي تواجه عملية تطوير المدقق الآلي. وسنستعرض في هذا الفصل طريقة بناء المعاجم مع الاعتبارات التي يمكن أن يضعها المطور خلال بنائه لها.

٢ , ٢ الأخطاء الإملائية الشائعة

تكون الأخطاء الإملائية الناتجة عن الكتابة بالحاسوب على شقين: إما أن تكون أخطاء إدراكية (Cognitive Errors)، و هي تلك التي تنتج عن عدم معرفة بالإملاء الصحيح للكلمة، وهذا النوع من الأخطاء يكون مشتركا في النصوص المكتوبة بالحاسوب أو تلك المكتوبة باليد، أو أن تكون أخطاء طباعية (Typographical Errors)، و هي تلك التي تنتج عن حدوث خلل أثناء إدخال الكلمات بواسطة لوحة المفاتيح.

بداية فإن الأخطاء الإدراكية في اللغة العربية تقع في الغالب من:

- أحرف لها نفس الصوت. ومثال ذلك التاء المربوطة والتاء المفتوحة:

الصواب: فعاليات	تكتب خطأ: فعالياة
الصواب: قضاة	تكتب خطأ: قضات

- أحرف لها نفس الصوت عند الوقف أو الابتداء. ومثال ذلك همزتي الوصل والقطع، والتاء المربوطة والهاء:

الصواب: ابن	تكتب خطأ: إبن

الصواب: أستعمل (للمضارع)	تكتب خطأ: استعمل
الصواب: حديقة	تكتب خطأ: حديقه
الصواب: أرداه	تكتب خطأ: أرداة

- أحرف تنطق ولا تكتب. مثال ذلك الألف اللينة:

الصواب: هذا	تكتب خطأ: هاذا
الصواب: لكن	تكتب خطأ: لاكن

أما الشق الثاني من الأخطاء، وهو الأخطاء الناتجة عن خلل أثناء إدخال الكلمات عن طريق لوحة المفاتيح، فإنه يظهر عادة على شكل أربع حالات (Haddad و Yassen، 2007، Kukich، 1992) وهي:

١. حذف حرف، ومثال ذلك:

الكلمة المعنية: يستمر	**المدخلة**: يتمر	(حذف الحرف الثاني)

٢. إضافة حرف، ومثال ذلك:

الكلمة المعنية: يستمر	**المدخلة**: يسشتمر	إضافة حرف بين الحرفين (الثاني والثالث

٣. تبديل حرف بحرف آخر، ومثال ذلك:

الكلمة المعنية: يستمر	**المدخلة**: يشتمر	تبديل الحرف الثاني من) (س إلى ش

٤. قلب حرفين متجاورين، ومثال ذلك:

الكلمة المعنية: يستمر	**المدخلة**: يسمتر	(قلب الحرفين الثالث والرابع)

وقد أضافت Kukich صنفا ثالثا من الأخطاء وهو: الأخطاء الناتجة عن تشابه الأصوات (Phonetic Errors). ويمكن للأخطاء الناتجة عن تشابه الأصوات أن

تكون حالة خاصة من الأخطاء الإدراكية، إذ إن الخطأ في الإملاء الناتج عن تشابه الأصوات لبعض الحروف قد يكون ناتجا عن عدم معرفة الكاتب بالحرف الصحيح الموجود في الكلمة فيستبدله بحرف آخر له نفس الصوت (Kukich، 1992)، وقد يكون الخطأ ناتجا عن عدم التركيز، بالرغم من معرفة الكاتب بالإملاء الصحيح للكلمة وبالتالي يمكن أن يندرج هذا الخطأ تحت الأخطاء الطباعية.

أما الأخطاء الطباعية فتكون متوقعة وشائعة بسبب الرغبة في الإدخال السريع للكلمات من قبل الكاتب. فأثناء الإدخال السريع يحدث ألا يتم الكبس على زر أحد الأحرف، أو أن يتم الكبس على زر حرف إضافي وعندها تحدث الحالتان الأولى والثانية من الحالات الأربع أعلاه. وأما الحالة الثالثة فإنها قد تحدث بسبب قرب الحرفين الذين تم تغييرهما في الكلمة على لوحة المفاتيح مما يجعل المدخل يكبس على زر الحرف الخطأ بدلا من الحرف الصحيح. وهذه جميعها يمكن تصنيفها تحت الأخطاء الطباعية. ومن أسباب الحالة الثالثة أيضا تشابه صوتي الحرفين الذين تم تغير أحدهما بالآخر والذي يجعل المدخل يدخل الحرف الخطأ والذي له نفس صوت الحرف الصحيح كما هو الحال في كلمتي "فعاليات" و"فعالياة" مثلا. و هذا النوع من الأخطاء قد ينتج بسبب ضعف التركيز والرغبة في الإدخال السريع، أو الجهل بالإملاء الصحيح للكلمة. لذا يمكن تصنيف هذا النوع من الأخطاء إلى أخطاء ناتجة عن تشابه الأصوات، أو بصورة أعم إلى أخطاء إدراكية في حالة الاعتقاد بعدم معرفة الكاتب بالإملاء الصحيح للكلمة، أو إلى أخطاء طباعية في حالة الاعتقاد بمعرفة الكاتب بالإملاء الصحيح للكلمة. أما الحالة الرابعة فإنها تحدث عادة لتسابق إصبعي المدخل أثناء الإدخال السريع فيدخل حرفا قبل الآخر بشكل خاطئ. و هذه الحالة يمكن أن تندرج تحت الاخطاء الطباعية.

وقد تحدث الأخطاء الطباعية بوعي من المدخل وذلك لأسباب منها أن بعض الأحرف تحتاج إلى أن يقوم المدخل بالكبس على زرين اثنين بدلا من زر واحد، وتحدث هذه كثيرا في الحروف المهموزة. فمثلا الحرف "أ" يلزم أن يقوم المدخل بالضغط على الزر Shift ومن ثم زر الحرف "ا". ونظرا لأن المدخل يرغب في الإدخال السريع فإنه قد يلجأ للضغط مباشرة على الزر "ا" لكي يتفادى الضغط على زرين ولأن الحرفين متشابهان في النطق والكتابة (Buckwalter، 2004). وأيضا قد يحدث الإدخال الخطأ نتيجة لعدم وجود الحرف المطلوب في لوحة المفاتيح. وهذا النوع من الأخطاء يحدث

كثيرا في النصوص المدخلة في بداية تعريب الحاسوب، إذ إن لوحات المفاتيح المصنعة في ذلك الوقت كانت تحتوي على الحروف الهجائية الرئيسة وبعض مشتقاتها الأساسية فقط. فمثلا، الحروف "ؤ"، "ئ"، "آ"، و"لآ" ربما لا توجد في بعض لوحات المفاتيح القديمة. وبالتالي يلجأ المدخل لاستخدام الحروف الهجائية الرئيسة المقابلة لها. فمثلا يتم إدخال كلمة "الآخرة" بهذا الشكل "الاخرة" أو بهذا الشكل "الأخرة". وأيضا كلمة "المسؤول" يمكن أن يتم إدخالها بهذا الشكل "المسوول".

واستعراضنا لطبيعة الأخطاء الإملائية وكيفية حدوثها في هذا الجزء يأتي من رغبتنا في إعطاء القارئ الكريم نبذة عن بعض مسببات هذه الأخطاء. وبالإمكان أخذ هذه المسببات بعين الاعتبار في تطوير المدققات الإملائية سواء من حيث الاكتشاف أو التصحيح.

٢, ٣ المدقق الإملائي

يعتبر اكتشاف الأخطاء الإملائية (Spelling Error Detection) المكون الأساس للمدقق الإملائي. ومن ثم فإن تصحيح الخطأ (Spelling Error Correction) يعتبر عنصرا إضافيا. لذا فإنه بالإمكان تطوير مدقق إملائي يتكون من مكتشف الأخطاء فقط دون أن يقوم باقتراح الكلمة الصحيحة. وفي هذه الحالة يقوم المدقق الإملائي بالإشارة إلى الكلمات الخطأ في النص وإبرازها للكاتب وترك الكاتب ليقوم بتعديلها دون إعطائه أية مقترحات. ولكن المدققات الإملائية الحديثة تعتمد الاكتشاف والتصحيح، إذ إن تصحيح الخطأ هو عبارة عن اقتراح بعض الكلمات للكاتب والتي قد تكون إحداهن هي الكلمة الصحيحة، أو أن يقوم المصحح باختيار إحدى الكلمات التي تعتبر الأقرب للصواب بناء على الخوارزمية المستخدمة ويضعها مكان الكلمة الخطأ بشكل آلي. وعملية تصحيح الخطأ تعتبر أكثر تحديا من عملية اكتشاف الخطأ، إذ إن معرفة الكلمة التي يريدها الكاتب عملية يحيط بها الكثير من الغموض.

أما بالنسبة لاكتشاف الأخطاء، فإنه بداية يمكن تصنيف الأخطاء إلى صنفين:

١. خطأٌ يُنتج كلمةً ليست من كلمات اللغة (Non-Word Error).

٢. خطأٌ يُنتج كلمةً من كلمات اللغة، ولكنها لا تتناسب مع السياق، وتخل بمعنى الجملة (Real-Word Error).

و مثال الصنف الأول:

وضعت الكأس **فقو** الطاولة

فبدلا من إدخال كلمة «فوق» تم إدخال كلمة «فقو» والتي ليست من كلمات اللغة العربية.

و مثال الصنف الثاني:

وضعت الكأس **وفق** الطاولة

فكلمة «وفق» من كلمات اللغة العربية، ولكنها ليست ذات معنى في سياق الجملة، وتعتبر الجملة بوجود هذه الكلمة جملة لا معنى لها.

ويعتبر اكتشاف الصنف الثاني من الأخطاء أكثر تحديا من اكتشاف الصنف الأول، إذ إنه لاكتشاف خطأ من هذا الصنف، يحتاج الشخص إلى ربط الكلمة بالكلمات الأخرى في الجملة ومن ثم معرفة السياق. وبحكم أن اكتشاف هذا النوع من الأخطاء يحتاج إلى بعض الخوارزميات المتقدمة، فإننا في هذا الفصل، والذي يهدف إلى إعطاء مقدمة عن المدققات الإملائية، سنتكلم عن اكتشاف الصنف الأول من الأخطاء فقط.

و فيما يتعلق بتصحيح الأخطاء، فإنه أيضا يمكن تصنيف طرق التصحيح الرئيسة إلى صنفين (Kukich، 1992):

١. تصحيح الكلمة بشكل مستقل (Isolated-Word Error Correction).

٢. تصحيح الكلمة مع أخذ السياق بعين الاعتبار (Context-Dependent Word Correction).

فالطريقة الأولى تقوم على أساس تصحيح الكلمة دون اعتبار للسياق عند اقتراح الكلمات الصحيحة. فلو أخذنا الجملة السابقة:

وضعت الكأس **فقو** الطاولة

فإن المصحح قد يقترح الكلمات التالية:

فقول، فوق، قو، فقوا، فقه، وفق، ...

على الرغم من أنه، باستثناء كلمة «فوق»، فإن الكلمات الأخرى المقترحة لا تناسب السياق، ولا تعطي للجملة معنى مفهوما، ولكن هذه الكلمات تم اقتراحها لأن المصحح لا يأخذ السياق بعين الاعتبار.

أما في الطريقة الثانية، فإن المصحح ربما يقترح كلمة «فوق» فقط دون الكلمات الأخرى التي تم اقتراحها في الطريقة الأولى؛ لأنه درس السياق ووجد أن هذه الكلمة هي الكلمة المناسبة لسياق الجملة. وفي هذا الفصل سنكتفي بدراسة الطريقة الأولى فقط.

٣. اكتشاف الأخطاء

يعتبر اكتشاف الخطأ هو الخطوة الأولى في التدقيق الإملائي، وربما تكون الخطوة الوحيدة في حالة أن المدقق الإملائي يكتفي باكتشاف الخطأ دون تصحيحه. وهنالك طريقتان شائعتان لاكتشاف الأخطاء. الطريقة الأولى، وهي الأكثر تطبيقا بحسب علم الكاتب، هي تلك التي تعتمد استخدام المعاجم اللغوية في اكتشاف الأخطاء. وتستخدم هذه الطريقة لاكتشاف الأخطاء بناء على الكلمات فقط دون النظر إلى السياق. والطريقة الثانية هي تلك التي تستخدم نظرية الاحتمالات، وتُستخدم هذه الطريقة عادة في اكتشاف الأخطاء بناء على السياق. وسنكتفي بمناقشة الطريقة الأولى في هذا الجزء.

تعتبر المعاجم اللغوية المصدر والمرجع الأساس لكلمات اللغة، وتتكون من مجموعة من كلمات اللغة واشتقاقاتها. وتختلف المعاجم في حجمها، فقد يتكون معجم ما لإحدى اللغات من عدد معين من الكلمات وقد يكون هنالك معجم آخر لنفس اللغة يحتوي على عدد أكبر من الكلمات. ويمكن ضرب مثال على اختلاف المعاجم بالحصيلة اللغوية للأشخاص. فالحصيلة اللغوية لشخص ما هي الكلمات المخزنة في ذاكرته والتي يمكن اعتبارها معجما ذهنيا للشخص. فقد يأتي شخص آخر يتكلم نفس اللغة ولكن بحصيلة لغوية مختلفة، أي بعدد كلمات مختلف وبمجموعة مختلفة عن المجموعة التي لدى الشخص الأول بحيث تتقاطع المجموعتان بعدد لا بأس به من الكلمات. فالمعجم اللغوي لشخص في سن السابعة مثلا، أقل بكثير في الظروف الاعتيادية للمعجم اللغوي لشخص في سن العشرين. وكذلك الحال بالنسبة للمعاجم المدونة، فقد يحتوي معجم ما تم بناؤه من ذخيرة لغوية معينة على مجموعة من الكلمات تختلف عن معجم تم بناؤه من ذخيرة لغوية أخرى. فلو أردنا على سبيل المثال بناء معجم لغوي من نصوص مأخوذة من كتب في الاقتصاد، فإن هذا المعجم سيختلف عن معجم آخر تم بناؤه من نصوص مأخوذة من كتب في الطب. بل إنه لو بنينا معجمين لغويين من كتب في نفس الفن ولكن لمؤلفين مختلفين بحيث يكون كل معجم مبني من كتب أحد المؤلفين، لربما

خرجنا بمعجمين مختلفين أيضا، لأن الحصيلة اللغوية لكل مؤلف ستنعكس على كتابته واستخدامه للكلمات.

إن اكتشاف الأخطاء الإملائية بالرجوع إلى المعاجم اللغوية من دون النظر إلى السياق هي عملية لا تتعدى البحث في المعجم عن كل كلمة في النص المعالج. فإذا كانت الكلمة موجودة في المعجم فإن هذه الكلمة تعتبر صحيحة، أما إذا كانت الكلمة غير موجودة في المعجم فإنها تعتبر خطأ. لذا فإنه من الضروري في تصميم المعاجم التي تستخدم في المدققات الإملائية أخذ النقاش السابق بعين الاعتبار. فإن اتخاذ قرار بعدم إضافة كلمة إلى المعجم يعني أن المدقق الإملائي سيعتبر هذه الكلمة خطأ عند المرور بها في أي نص، وكذلك قرار إضافة كلمة إلى المعجم يعني أن هذه الكلمة ستعتبر من كلمات اللغة، وبالتالي فستعتبر صحيحة. وبالنظر إلى هذا فإنه قد يتم بناء معاجم مخصصة لفن من الفنون، أي أنه قد يتم بناء معجم لغوي يستخدم في التدقيق الإملائي للكتابة في مجال الاقتصاد مثلا. وقد يبنى هذا المعجم من ذخيرة لغوية كبيرة مؤلفة من كتب ومقالات في الاقتصاد لمؤلفين كثر بحيث تحتوي على أكبر عدد من الكلمات في هذا الفن. وربما يحتوي هذا المعجم على بعض الكلمات التي في أصلها ليست من اللغة التي يكتب بها النص ولكنها مستوردة من لغات أخرى. فلو كان النص عربيا فقط سيحتوي المعجم على كلمات إنجليزية معربة تستخدم في مجال الاقتصاد.

لكن تصميم معجم لكل فن بحيث يستخدم الكاتب المعجم الذي يريد بناءً على موضوع كتابته قد لا يروق لكثير من المستخدمين. إذ إن المستخدم قد يكتب كتابا أو مقالة عامة تتطرق لأكثر من فن في آن واحد، وفي كثير من الأحيان قد يكتب المؤلف موضوعا في فن ما ويستلهم نقاطا من فنون أخرى؛ لذا فإن الحاجة كبيرة لمعجم عام للغة يمكن استخدامه للتدقيق الإملائي، هذا المعجم يمكن بناؤه من ذخيرة لغوية مكونة من نصوص مأخوذة من فنون ومعارف شتى، وربما تكون هذه النصوص من حقب زمنية مختلفة ومكتوبة بواسطة مؤلفين من أقطار مختلفة، والمعجم الناتج في هذه الحالة ربما يكون كبيرا بالمقارنة مع المعاجم المتخصصة. لكن هنالك إشكالية في هذا النوع من المعاجم أيضا، فنظرا لأنه يحتوي على عدد كبير من الكلمات، فإن كثيرا من الكلمات قد لا تستخدم إلا بنسب قليلة جدا في النصوص التي تكتب، وهذه الكلمات تشكل مشكلة للمدقق الإملائي، إذا إن الغالبية من الكُتَّاب قد لا يحتاجون هذه الكلمات،

ولكنهم يستخدمون كلمات أخرى مشابهة لها في الإملاء. وفي حالة وقوعهم في أخطاء في كتابة الكلمات التي يريدون، ونتج عن هذه الأخطاء هذه الكلمات القليلة الاستخدام، فإن المدقق الإملائي لن يعتبرها خطأً؛ نظرا لوجودها في المعجم (Peterson، 1980). فعلا سبيل المثال، لو أخذنا كلمة "الكرى" وتعني النوم أو النعاس، فهذه الكلمة ربما قل استخدامها في النصوص، فقد يقول قائل إن إضافتها للمعجم المستخدم بواسطة المدقق الإملائي قد تمنع المدقق الإملائي من اكتشاف أخطاء لكلمات أخرى مشابهة، كتبت خطأً على هيئة هذه الكلمة مثل "الثرى" و"الكرم". فلو أن الكاتب أخطأ في كتابة "الكرم" وكتبها "الكرى" فإن المدقق الإملائي لن يشير إلى كلمة "الكرى" على أنها خطأ نظرا لوجودها في المعجم، وفي المقابل فإن آخرين قد يرون إضافة جميع الكلمات المعروفة إلى المعجم الخاص بالمدقق الإملائي. وعلى أية حال، فإنه لا توجد قاعدة معينة يمكن تطبيقها على ما هي الكلمة التي يجب أن تضاف إلى المعجم وما هي الكلمة التي يجب ألا تضاف، ويبقى هذا القرار خاضعا لمصمم المدقق الإملائي. ويذكر Peterson أنه في حالة معالجة ذخيرة لغوية لبناء معجم للمدقق الإملائي، فإنه بالإمكان وضع حد معين بحيث إن أي كلمة تتكرر في النص بأقل من هذا الحد تعتبر كلمةً قليلة الاستخدام، ولا تضاف للمعجم حتى لا تُشكِل على غالبية المؤلفين، فيما تضاف جميع الكلمات التي تزيد عن هذا الحد (Peterson، 1980). أما تحديد هذا الحد فهو قرار هندسي يخضع للمصمم، وفي حال لو وردت كلمة ما في النص واعتبرها المدقق الإملائي خطأً بينما هي صحيحة فإن الكثير من المدققات الإملائية - كالمدقق الإملائي لمحرر مايكروسوفت (Microsoft Word) والمدقق الإملائي آسبل[1] (Aspell) - تتيح للمستخدم إمكانية إضافة هذه الكلمة إلى المعجم اللغوي، فتصبح من كلمات اللغة التي يستخدمها المؤلف. فلو افترضنا أن كلمة "الكرى" في المثال أعلاه ليست موجودة في المعجم اللغوي للمدقق الإملائي، وقام المؤلف باستخدامها وهو يعنيها، فإن المدقق الإملائي بعد أن يشير إلى أن هذه الكلمة خطأ، فإنه سيعطي الخيار "إضافتها إلى المعجم» إلى المؤلف. وفي حالة أن المؤلف اختار هذا الخيار، فإن الكلمة ستكون من مجموع كلمات المعجم، ولن يشير إليها المدقق الإملائي ككلمة خطأ عند استخدامها مرة أخرى.

1- http://aspell.net/

وعادة ما يقوم مصمم المدقق الاملائي باتخاذ قرارته الهندسية بناء على هذه المعايير:

a) عدد الكلمات الخطأ التي تم التعرف عليها (Number of True Positive Cases). وسيتم الإشارة إلى هذا المعيار اختصارا بـ TP.

b) عدد الكلمات الخطأ التي لم يتم التعرف عليها (Number of False Negative Cases). وسيتم الإشارة إلى هذا المعيار اختصارا بـ FN.

c) عدد الكلمات الصحيحة التي تمت الإشارة إليها على أنها خطأ (Number of False Positive Cases). وسيتم الإشارة إلى هذا المعيار اختصارا بـFP.

فإنه عند تصميم المدقق الإملائي، تأتي الرغبة دائما في زيادة المعيار الأول وتقليل المعيارين الثاني والثالث. وبناء على هذه المعايير الثلاثة يأتي تقييم أداء المدقق الإملائي بحساب ثلاث قيم رياضية:

الدقة (Precision) $= \dfrac{TP}{TP + FP}$

وتمثل نسبة الكلمات الخطأ التي تم التعرف عليها بشكل صحيح من بين الكلمات التي تمت الإشارة إليها على أنها خطأ.

الاسترجاع (Recall) $= \dfrac{TP}{TP + FN}$

وتمثل نسبة الكلمات الخطأ التي تم التعرف عليها بشكل صحيح من بين الكلمات الخطأ الموجودة في النص.

مقياس إف (F-measure) $= 2 \times \dfrac{Precision \times Recall}{Precision + Recall}$

و يمثل معدل الدقة والاسترجاع.

فلو افترضنا أنك أردت أن تقيم أداء أحد المدققات الإملائية وأدخلت له نصا فيه ٥٠٠٠ كلمة، ٢١٣ منها خطأ. فلو اكتشف المدقق الإملائي ١٩٠ كلمة من هذه الـ ٢١٣، بينما أشار إلى ١٥ كلمة صحيحة على أنها خطأ، فإن أداء هذا المدقق الإملائي سيكون كالتالي:

TP = 190 FP = 15 FN = 213 – 190 = 23

F-measure = 90.7 % Recall = 89.2 % Precision = 92.7 %

٤. تصحيح الأخطاء

يمكن شرح عملية التصحيح الآلي للأخطاء الإملائية بأنها إيجاد الكلمات الأقرب للصواب لتحل محل الكلمات الخطأ، ومن هذا المنطلق العام، فإن كل مطور بإمكانه أنه يعمل على تطوير خوارزميته الخاصة لإيجاد الكلمات الأقرب للصواب بناء على اعتباراته. وسنشرح في هذا الجزء إحدى الطرق الأساسية المطورة في مجال التصحيح والتي تعتمد على الكلمة الخطأ فقط، كما ذكرنا سلفا، دون أخذ السياق بعين الاعتبار.

بما أن الكلمة الخطأ هي المعلومة الوحيدة التي يمكن عن طريقها استنتاج الكلمة الصحيحة، فإن أحد مبادئ التصحيح الشائعة يقوم على أساس أن هذه الكلمة هي إحدى الكلمات الصحيحة ولكنها تعرضت لبعض التغيير نتيجة لخطأ ما. لذا فإنه يمكن استنتاج الكلمة الصحيحة من هذه الكلمة الخطأ، وذلك بافتراض أن الكلمة الصحيحة هي إحدى الكلمات القريبة هجائيا من هذه الكلمة. وبالحديث عن القرب، فإنه لابد من تعريف «**المسافة**» والتي يمكن من خلالها تحديد قرب كلمتين من بعضهما. ولأننا نرغب في معرفة قرب كلمتين من بعضهما هجائيا فإن المسافة التي نريد تعريفها هنا يجب أن تكون مسافة هجائية.

يمكن لأي مطور لمدقق إملائي أن يعرف المسافة بالطريقة التي يراها تحسن من أداء التصحيح [١]. ومن التعريفات التي تم وضعها للمسافة بين كلمتين والتي تستخدم كثيرا في معالجة اللغات البشرية وفي التصحيح الإملائي هي مسافة دميراو–ليفينشتاين (Damerau–Levenshtein) و التي يمكن تعريفها بالآتي (Damerau، 1964):

"**مسافة دميراو–ليفينشتاين** بين كلمتين: (م) ويطلق عليها الكلمة المصدر، و(هـ) ويطلق عليها الكلمة الهدف، هي أقل عدد من العمليات التالية: إضافة حرف، حذف حرف، تبديل حرف بحرف آخر، أو قلب حرفين متجاورين، والتي يمكن إجراؤها على الكلمة (م) لتحويلها إلى الكلمة (هـ)".

١– وقد يعتمد طرقا أخرى للتصحيح غير المسافة.

ويطلق على هذه المسافة أيضا في بعض المراجع **بمسافة أقل العمليات التحريرية** (Minimum Edit Distance). والعمليات الأربع المذكورة في التعريف هي العمليات التي تم شرحها في الجزء ١,٢. ولكي نأخذ مثالا على هذه المسافة، فلنأخذ الكلمة "فقو" كما في المثال المذكور في الجزء ١,٣ ونعتبرها هي الكلمة المصدر أو الكلمة (م) كما في التعريف. فإن جدول ١ أدناه يعطي بعض الكلمات الهدف والمسافة بينها وبين هذه الكلمة.

الكلمة المصدر (م): فقو		
الكلمة الهدف (هـ)	**المسافة**	**العمليات**
فقول	١	إضافة حرف في أخر الكلمة.
فوق	١	قلب الحرفين الأخيرين.
وفق	٢	قلب الحرفين الأخيرين فتصبح الكلمة «فوق». ومن ثم قلب الحرفين الأولين للكلمة الجديد "فوق" فتصبح "وفق".
فور	٢	حذف الحرف الثاني. إضافة حرف في نهاية الكلمة.
واق	٣	تبديل الحرف الأول «ف» بـ «و». تبديل الحرف الثاني "ق" بـ "ا". تبديل الحرف الثالث "و" بـ "ق" وهنالك طريقة أخرى وهي: حذف الحرف الأول فتنتج الكلمة "قو". قلب الحرفين للكلمة الجديدة فتنتج الكلمة "وق" إضافة حرف "ا" بين الحرفين الأول والثاني للكلمة الناتجة من العملية الثانية.

الكلمة المصدر (م): فقو		
الكلمة الهدف (هـ)	**المسافة**	**العمليات**
يرد	٣	تبديل الحرف الأول «ف» بـ «ي». تبديل الحرف الثاني "ق" بـ "ر". تبديل الحرف الثالث "و" ب "د"

جدول ١ : مسافة دميراو-ليفينشتاين بين الكلمة "فقو" وبعض الكلمات الأخرى.

فتطوير مصحح إملائي بالاعتماد على هذه المسافة يكون باعتبار أن الكلمة المصدر هي الكلمة الخطأ، والبحث في المعجم عن الكلمات التي تبعد مسافة معينة عن هذه الكلمة، ومن ثم اقتراحها للمستخدم. يمكن تحديد سقف أعلى للمسافة التي تؤخذ بعين الاعتبار للبحث عن الكلمات المقترحة حتى لا يتم اقتراح قائمة طويلة جدا من الكلمات. فمثلا يمكن تصميم المصحح الإملائي بحيث يقترح الكلمات التي تبعد ٢ فقط بحد أعلى عن الكلمة الخطأ. وعلى هذا الأساس يمكن ترتيب الكلمات المقترحة من حيث قربها للصواب بحسب مسافتها من الكلمة الخطأ. فمثلا توضع الكلمات التي تبعد المسافة ١ عن الكلمة الخطأ في أعلى قائمة الكلمات المقترحة باعتبار أنها ربما تكون هي الأقرب للصواب. ومن ثم توضع الكلمات التي تبعد المسافة ٢، ومن ثم المسافة ٣، إذا كان المصحح يذهب أبعد من ٢، وهكذا إلى أن يصل إلى الكلمات التي تبعد الحد الأعلى الذي تم تحديده وتكون هذه الكلمات في مؤخرة القائمة.

٥. إشارة ختامية إلى الفكرة العامة لبعض التقنيات المتقدمة وبعض المراجع

العديد من الطرق المتقدمة والتي تم تصميمها للتدقيق الإملائي تعتمد على نظرية الاحتمالات (Probability Theory) وخاصة في عمليتي التصحيح بشكلها العام، واكتشاف الأخطاء بناء على السياق وليس بناء على الكلمة فقط. ونظرية الاحتمالات هي النظرية الرياضية التي تعطي قياسا لعدم التأكد (Uncertainty)، أو درجة الاعتقاد (Degree of Belief). فلو سألت شخصا عن صحة حادثة معينة، وهو غير متأكد من صحتها، فقد يجيبك بـــــ "أعتقد أنها صحيحة بنسبة كذا". وهذه النسبة تمثل درجة

اعتقاده بصحة الحادثة. فعلى سبيل المثال، لو أن شخصا تحدث وقال "السلام" ثم توقف لبرهة، فإن شخصا آخر ربما يقول "أعتقد بنسبة ٩٠ في المائة أن الكلمة التالية هي "عليكم"، وهذا يمثل عدم تأكده الكامل مما سيقوله المتكلم، إذ إنه ربما يقول "على" يريد بها "على الحضور" مثلا. وكذلك يمكن تصميم مدقق إملائي بحيث يكتشف ويصحح الأخطاء بناء على نموذج احتمالي للغة يمكن من خلاله قياس احتمال صواب أو خطأ الكلمات وفق سياقاتها. إذ إنه يمكن من خلال هذا النموذج اكتشاف وتصحيح الأخطاء بناء على درجة الاعتقاد بالصواب والخطأ.

وأضع هنا بعض المراجع التي يمكن من خلالها الاستزادة من هذه المواضيع، وأعتذر للقارئ الكريم من أن هذه المراجع جميعها باللغة الإنجليزية. يمكن الرجوع لـ Casella و Berger (Casella و Berger ، 2002) للتوسع في نظريات الاستنتاج الإحصائي والاحتمالات. أما بالنسبة لاكتشاف الأخطاء وتصحيحها باستخدام النماذج الاحتمالية فإنه يمكن الرجوع لـــلفصل الخامس من Martin و Jurafsky) Jurafsky و Martin 2000،). ويمكن الرجوع لـ Kernighan وآخرين (Kernighan وآخرون، ١٩٩٠) وكذلك Brill و Moore) Moore و Brill ، 2000) للاطلاع على كيفية استخدام نموذج احتمالي مستوحى من نظرية المعلومات والاتصالات يعرف **بالقناة الصاخبة** (Noisy Channel) لتصحيح الأخطاء الإملائية. وهذا النموذج يستخدم كثيرا في الأبحاث وربما بعض الأنظمة الحديثة في التصحيح الإملائي. ويمكن أيضا الرجوع لـ Church) Gale Church و Gale ، 1991) للاستزادة من كيفية استخدام نظرية الاحتمالات لتصحيح الأخطاء الإملائية.

شكر وإهداء

الحمد لله أولا وقبل كل شيء، فله الفضل والمنة على ما يسر من سبل الحصول على المعرفة ومن تيسير إتمام هذا العمل.

أشكر أسرة تحرير الكتاب على دعوتهم لي لكتابة هذا الفصل، وأتمنى أن أكون قد وفقت في كتابة ما يفيد القارئ الكريم.

الكثير مما قرأته ومن ثم كتبته في هذا الفصل تعلمته أثناء عملي في أبحاث لتقويم المقالات العربية وكذلك التدقيق الإملائي للغة العربية. وممن شاركت معهم في هذه الأعمال واستنرت بأفكارهم واستفدت من نقاشاتهم: الدكتور محمد الكنهل، عبدالعزيز القباني، محمد الحمادي، علي عريشي، أثير الخليفة، ولمياء القويعي. كما أشكر الأستاذ منتصر أحمد الذي قام بمراجعة لغوية لهذا الفصل، والشكر موصول أيضا للدكتور عبدالفتاح عبيد. ويجب التنويه على أن أي خطأ في هذا الفصل فهو مني وحدي.

والداي الكريمان، الدكتور عبدالله الصانع و سارة البطي، لهما الفضل بعد الله سبحانه وتعالى فيما تعلمته وعملته، أسأل الله سبحانه وتعالى أن يجزيهم عني خير الجزاء.

دائما وأبدا أشكر زوجتي نوف الرويشد على صبرها على انشغالي وغيابي الذهني الكثير خلال قراءاتي وعملي، وأسأل الله سبحانه وتعالى أن يجعل ذلك في موازين حسناتها.

أهدي هذا الجهد المتواضع لأبنائي عبدالعزيز، نواف، وسارة.

المراجع

◇ **Turing, Alan M**. Computing machinery and intelligence. Mind, 59, 433-460.

◇ **Russell, Stuart and Norving, Peter.** Artificial Intelligence: A Modern Approach. Prentice Hall, United States, International Edition, 1995.

◇ **Damerau, Fred J**. A Technique for Computer Detection and Correction of Spelling Errors. Communications of the ACM, vol. 7, 3, 1964, 171-176.

◇ **Haddad, Bassam and Mustafa, Yaseen**. Detection and Correction of Non-Words in Arabic: A Hybrid Approach. Internation Journal of Computer Processing of Language, vol. 20, 4, 2007.

◇ **Kukich, Karen**. Techniques for Automatically Correcting Words in Text. ACM Computing Surveys, vol. 24, 4, 1992, 377-439.

◇ **Buckwalter, Tim**. Issues in Arabic Orthography and Morphology Analysis. Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages (Semitic'04), 2004, 31-34.

◇ **Peterson, James L**. Computer Programs for Detecting and Correcting Spelling Errors. Communications of the ACM, vol.23, 12, 1980, 676-687.

◇ **Casella, George and Berger, Roger L**. Statistical Inference. Cengage Learning, New Delhi, India, 2nd Edition, 2002.

◇ **Jurafsky, Daniel and Martin, James H**. Speech and Language Processing. Prentice Hall, New Jersey, United States, 1st Edition, 2000.

◇ **Kernighan, Mark D., Church, Kenneth W. and Gale, William A**. A Spelling Correction Program Based on a Noisy Channel Model.

Proceedings of the 13th conference on Computational linguistics (COLING'90), vol.2, 1990, 205-210.

◇ **Brill, Eric and Moore, Robert C**. An Improved Error Model for Noisy Channel Spelling Correction. Proceedings of the 38th Annual Meeting on Association for Computational Linguistics (ACL'00), 2000, 286-293.

◇ **Church, Kenneth W. and Gale, William A**. Probability Scoring for Spelling Correction. Statistics and Computing, Springer, vol.1, 2, 1990, 93-103

◇◇◇

فهرس الفصول

فهرس المحتويات

هذه الطبعة
إهداء من المركز
ولا يسمح بنشرها ورقياً
أو تداولها تجارياً

مباحث لغوية ٣٠

مدخل إلى اللسانيات الحاسوبية

يُصدِر مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية هذا الكتاب ضمن سلسلة (مباحث لغوية)، وذلك وفق خطة عمل مقسمة إلى مراحل، لموضوعات علمية رأى المركز حاجة المكتبة اللغوية العربية إليها، أو إلى بدء النشاط البحثي فيها، واجتهد في استكتاب نخبة من المحررين والمؤلفين للنهوض بعنوانات هذه السلسلة على أكمل وجه.

ويهدف المركز من وراء ذلك إلى تنشيط العمل في المجالات التي تُنَبّه إليها هذه السلسلة، سواء أكان العمل علميا بحثيا، أم عمليا تنفيذيا، ويدعو المركز الباحثين كافة من أنحاء العالم إلى المساهمة في هذه السلسة.

وتودّ الأمانة العامة أن تشيد بجهد السادة المؤلفين، وجهد محرر الكتاب، على ما تفضلوا به من رؤى وأفكار لخدمة العربية في هذا السياق البحثي.

والشكر والتقدير الوافر لمعالي وزير التعليم المشرف العام على المركز، الذي يحث على كل ما من شأنه تثبيت الهوية اللغوية العربية، وتمتينها، وفق رؤية استشرافية محققة لتوجيهات قيادتنا الحكيمة.

والدعوة موجّهة إلى جميع المختصين والمهتمين بتكثيف الجهود والتكامل نحو تمكين لغتنا العربية، وتحقيق وجودها السامي في مجالات الحياة.

الأمين العام
د.عبدالله بن صالح الوشمي

مركز الملك عبدالله بن عبدالعزيز الدولي
لخدمة اللغة العربية
King Abdullah Bin Abdulaziz Int'l Center for
The Arabic Language

ص.ب ١٢٥٠٠ الرياض ١١٤٧٣
هاتف:٠٠٩٦٦١١٢٥٨٧٢٦٨ - ٠٠٩٦٦١١٢٥٨١٠٨٢
البريد الإليكتروني: nashr@kaica.org.sa

9786039089520